METHODS OF STATISTICAL ANALYSIS

METHODS OF STATISTICAL ANALYSIS IN THE SOCIAL SCIENCES

BY

GEORGE R. DAVIES, PH.D.
Professor of Statistics, College of Commerce, University of Iowa

AND

WALTER F. CROWDER, PH.D.
Instructor in Commerce, College of Commerce, University of Iowa

NEW YORK
JOHN WILEY & SONS, INC.
LONDON: CHAPMAN & HALL, LIMITED
1933

PRINTED IN U. S. A.

PRESS OF
BRAUNWORTH & CO., INC.
BOOK MANUFACTURERS
BROOKLYN, NEW YORK

PREFACE

During recent years statistical methods have developed so rapidly in the social sciences—particularly in their business and civic uses—that introductory courses in the subject are necessarily becoming more highly specialized than formerly. The older text-books have generally given something of the philosophy of statistics, its field of application, the sources of data, and the various methods of charting, together with some of the elementary processes of mathematical analysis. But today a text-book which attempts to cover at all adequately, even for introductory purposes, all of these various aspects of the subject would either have to be encyclopedic in nature or would fail to do justice to most of them. Hence, in teaching statistics it is becoming necessary to place the emphasis upon those aspects of the subject which are most significant with respect to application in the fields to be studied, and to slight those elementary and specific phases with which the student is already somewhat familiar.

At the present time the tendency in an introductory course of statistics as applied to the social sciences is a concentration upon the methods and logic of statistical analysis. This concentration is justified because sources of data and fields of application are now more adequately considered in connection with many other courses making use of statistical data and affording opportunity for the extension of statistical research. Methods of presenting data in graphic form are somewhat familiar to the student, hence the introductory course need do little more than to summarize and illustrate the usual rules, leaving more specialized methods to be worked out in research practice by reference to the excellent manuals on charting now available. Such a reduction of the scope of the work makes it possible to concentrate attention upon laboratory methods and the logic underlying them, and to give sufficient practice in computation to fix these methods in mind.

This text-book has been prepared in view of the above considerations. The general scope of the subject is suggested in an introductory chapter, and specific reference to the logic of statistics is made in appropriate places throughout the book. Although references are made to sources of statistics, little is said specifically upon the subject of gathering data for the reason that this process can be learned only by practice

in a given applied field. The same thing may be said of specific fields of application. These fields are by no means standardized, and the research student who has learned a method and its significance must discover for himself in his chosen field of research just what use he can make of his tools. The exposition, however, of standard methods of computing averages, dispersion, index numbers, trends, cycles, and correlations has been expanded to include typical methods and procedures now in use, and to suggest ways of adapting such methods to specific problems. There is also included an abundance of simple exercises for laboratory practice. These exercises may be criticized in that many of them are abstract, and based on inadequate data, but it is hoped that these shortcomings may be more than offset by the emphasis which is given specific processes. They may, of course, be readily supplemented by applied laboratory work in such fields as may be determined by the student's primary interests.

In accordance with recommendations recently made by representative committees on the mathematical training of workers in the social sciences, contacts are made with several fields of mathematics. For example, simple applications of derivatives and integrations are suggested. Nevertheless, the stress has uniformly been placed upon application, and, as far as possible, all methods have been reduced to formulas so that they may be handled by the worker who has but a minimum of mathematical knowledge. The more important mathematical proofs are, however, suggested in footnotes or appendices so that the student who is mathematically inclined may be guided in a study of this aspect of the subject. Each chapter has been subdivided so as to give first the more elementary and basic processes, and later the more complex and specialized methods, thus facilitating the abbreviation of the course.

R. A. Fisher has well said that traditional statistics, built upon the theory of infinitely large samples, is inadequate for practical research. "Not only does it take a cannon to shoot a sparrow, but it misses the sparrow." Experience in laboratory work with irregular and scanty data demonstrates the advantages for many purposes of the simpler "first moment" methods over the methods involving higher powers, and it is for this reason that average deviation, quartiles, and first moment correlations have been stressed. It is the hope of the authors that the methods and procedure here outlined may be found useful to those who are now engaged in the worthy task of placing sociology, political science, and economics on a more adequate scientific foundation.

The authors wish to acknowledge their special indebtedness to Dr. Allen T. Craig, of the Department of Mathematics, University of

Iowa, and to Dr. Floyd B. Haworth, Mr. T. H. Cox, Mrs. Ruth McGuire, and Miss Gladys Hamilton, all of whom are affiliated with the College of Commerce, University of Iowa, for valuable assistance in preparing the manuscript.

George R. Davies,
Walter F. Crowder.

CONTENTS

METHODS OF STATISTICAL ANALYSIS IN THE SOCIAL SCIENCES

CHAPTER I

STATISTICAL METHODS IN SOCIAL SCIENCE

Except for sporadic beginnings, the use of statistical methods in the sciences of economics and sociology dates back scarcely more than a generation, and has seen the major portion of its development within the last decade. Some have been inclined to consider that this development has nearly run its course, but those who are in touch with the progressive methods now being used in large scale business and in governmental bureaus are of the opinion that statistics as applied to economic and social data is still at its beginning. Whether this is the case or not depends, however, upon the course of events. If, as seems probable, our present uncoordinated large scale business is to be further developed into an efficiently managed instrument of production serving the needs of the people, then statistics, together with mathematical economics, will emerge among the most important tools of the social sciences. For it is by means of averages, dispersions, coefficients of variability, trends, and regressions, as pictured in control charts, that management is able to visualize and direct the movements of large masses of population.

To be sure, figures may mislead us. They do not tell the whole truth. The work of the statistician is much like that of the map maker who presents the traveler with a sketch of important highways, showing the locations of towns and geographical features. The map is not a picture of reality. It shows cities as dots, and rivers as lines. It has purposely omitted the interesting details of scenery and the still more important features of human interest which lie along the route and which constitute the traveler's real objectives. Nevertheless, as a means of reaching these objectives, the map is extremely useful. And so it is with statistics in the hands of the business executive and the statesman. Back of the charts are human beings with their varying characteristics

and vital interests, few of which can be described in figures. Yet as a means of serving these interests, of keeping trade moving from one region to another, of allocating investment and labor, and of apportioning relief to maladjusted industries and dependent classes, statistics and mathematical methods are important, and are becoming increasingly important with the growing complexity of society.

The chief reason, then, for the study of statistics is the practical use which can be made of the subject in business and governmental control. That the work of the business executive and statesman will be considerably extended in the near future appears certain when we consider the chief social problems of the present time. Some of these problems arise out of business depressions. In past decades statistical procedure was applied to the task of describing business cycles as they occurred, and of tracing the interrelations of the factors concerned. Although it cannot be said that these studies have supplied a very clear theory of the causes of the business cycle, one fundamental fact has been distinctly revealed; namely, that depressions are related to lack of coordination of the various factors entering into a system of markets. This view is fully supported by mathematical economics, which shows that if the various industries and price levels are in suitable adjustment, one to another, there should result maximum production, a steady flow of goods to the consumer, and a distribution of income in proportion to the productivity of the various factors involved. A comparison of the theoretical picture of coordinated industry with statistical records of the business world shows how far the actual is removed from the ideal.

It is further evident that as industry becomes more complex the automatic adjustment of the various production and price schedules through the so-called law of supply and demand becomes more difficult, and intelligent supervision becomes increasingly necessary. Hence we may expect that the development of centralized control, such as has already been partially realized in national banking systems, industrial mergers, and similar organizations, will continue beyond its present scope. The tools which such centralized control will use will be chiefly statistics and accounting, together with such mathematical economics as may be required to meet problems of price equilibrium as they arise.

The development of a scientific basis of social control will require not only economic statistics but also social statistics, including data on standards of living, marriage and divorce rates, birth and death rates, dependency and crime, migration movements, and other demographic data. Indeed, it may be anticipated that all pertinent facts required in making a statistical map of the status and interrelations of a population will be brought together by centralized statistical bureaus much

more completely than they are today. It is only through such knowledge, and through intelligent business and political administration based upon it, that society can hope to meet the problems now facing it. The student of today, ambitious to reach the administrative positions of the future, will do well to master statistics at least to the degree requisite for intelligent criticism and utilization of its findings.

A less utilitarian reason for the study of statistics lies in the fact that statistical reasoning is essentially the logic of modern science, replacing the formal logic of the ancients. As a tool of logic, statistical reasoning focuses in the so-called contingency table, or double frequency distribution, such as is used in correlation. Physical scientists tell us that natural laws do not possess the fixity that formerly was attributed to them, but that they assume a certain degree of variability which in some cases may be appreciably large, and in other cases infinitesimally small. In the phenomena of industry and society, the variability of social "laws" or uniformities is large, but the mathematical form is essentially the same as for physical laws. Thus a correlation chart exhibiting the relationship of the price of corn to the production of corn, year by year, differs in degree of variability but not in principle from a physical law. And when inquiry is carried further, other independent variables are discovered which may be combined into a multiple correlation, improving the statement of the law and reducing the variability about the regression line.

From such a beginning one may proceed by the use of statistical and mathematical methods into the many complexities of economic law where any one price is, to some extent, a function of all other price and production schedules. Such a study, however, does not describe what actually occurs, but rather what would occur if men were intelligent enough to direct industry into the channels of maximum efficiency relative to the prevailing set-up of demand schedules and the present means of production.

By way of summary it may be said that the study of statistics is not merely an attempt to describe what actually occurs, though it must begin at this point, but in its broader aspects it is the logical background of business and social management. Hence what appears now to be mere abstraction may later become the basic necessity of an applied science. Eventually, it may be assumed, the social arts of business and politics will rest upon as substantial a theoretical and mathematical background as physics, chemistry, and engineering.

CHAPTER II

GATHERING AND PRESENTING DATA

It is not possible within the limits of a text-book to make an adequate exposition of the methods to be used in gathering, editing, and tabulating data, and presenting them in summary or graphic form. So much depends upon the object of the study and the nature of the available data that generalizations are of little use. It is scarcely possible to do more than to repeat certain obvious platitudes, such as that the object of the investigation should first be clearly defined, and that great care should be taken in the selection of the data, which must be consistent, homogeneous, reasonably reliable, ~~and~~ adequate. A little experience, however, will show that many difficulties are hidden beneath these apparently simple directions. To begin with, statistical units are often difficult to define. Thus in gathering figures on the number of farms in a given area, it may prove by no means easy to draw the line between the farm and the truck garden. Or in tabulating the number of dwelling houses, it may be a difficult task to discriminate among dwellings, apartment houses, and business buildings having apartments for rent. It is obvious that in such cases it will be necessary to set up precise and arbitrary definitions and to classify strictly in accordance with these definitions.

The consistency and homogeneity of figures will also challenge the student's powers of discrimination. Over a considerable period of time, such simple units as a bushel of wheat or a yard of cloth may vary considerably according to the quality listed. Or a series of items purporting to represent the interest rate may at one time be based upon three months' commercial paper and at another time on call-money loans—series which though distinctly related are not strictly comparable. If it is not possible to obtain strictly comparable units throughout the whole of a time series, it may be possible to make the data homogeneous by means of suitable adjustments. Such adjustments are, of course, very easily made when the units have common denominators, as, for example, when part of the data is listed in kilometers (0.62 mile) and part in miles, or in cases where the comparative money values of the diverse units listed are known. Sometimes very complex adjustments

may be required in editing the data for statistical analyses. Suppose, for example, that an index of bank debits in certain cities is being elaborated as a measure of the business cycle and that there are gaps in the data. In some cases such gaps may be filled in by reference to comparable data, such as bank clearings, after adjusting for the relative degree of variability, trend, and volume of the two series. But it is obvious that simple rules covering all such cases cannot be set up. Each case must be judged in accordance with principles covering the construction of indexes and the measurement of variability—topics which will be discussed later.

Sampling.—A large proportion of statistical analyses are made on the basis of samples rather than on the basis of complete data. It has been found by experience that, when samples are properly handled, they can be made to yield fairly satisfactory measures of the original data. It is a principle of science that large aggregates of data in which, from the point of view of sampling, the random element, or law of chance, is operative, exhibit a marked degree of statistical regularity, so that samples drawn from them tend to show the same characteristics as the aggregates themselves. For example, the tables by which life insurance premiums are calculated are based upon large aggregates of population, and are found by experience to be fairly accurate with respect to smaller groups of the same population. Similarly, small percentages, such as the relative number of exceptionally tall persons in a homogeneous population, are likely to remain stable from one random group to another. Thus in general the characteristics of a large group may usually be judged by studying a judicious sample collected at random.

In working with samples the question continually arises as to what constitutes an adequate number of items. This question is not easily answered, even if it is possible to obtain random samples without bias of any sort. It is evident, however, that as the size of the sample increases its dependability increases. Mathematically this principle is expressed somewhat more precisely by saying that dependability increases with the square root of the number of items; thus a sample of 100 items would normally be regarded as only twice ($\sqrt{100}/\sqrt{25}$) as dependable as one having 25 items. The term dependability as thus used may be rendered somewhat more precise by the following illustration. Let us assume that incoming freshman classes of very large numbers taken from year to year may be counted on to range normally in certain standardized mental tests through the following five classes of scores: 75–85, 85–95, 95–105, 105–115, 115–125, where 100 represents the average score. If successive groups each consisting of 25 students are selected at random from such a "universe" of students,

it would theoretically be expected that the averages of these groups would range in classifications one-fifth ($1/\sqrt{25}$) as broad as those of the original group. That is, these averages would range approximately in the classes 95–97, 97–99, 99–101, 101–103, 103–105. Again, if successive samples of 100 each are taken, the means of each of these samples might be expected to range in the classes 97.5–98.5, 98.5–99.5, 99.5–100.5, 100.5–101.5, 101.5–102.5. Thus, by increasing the size of the sample, the range through which the mean of the sample may be expected to vary approaches closer and closer to the mean of the original "universe." The theory of sampling of which the classifications just quoted represent a brief example constitutes an important branch of statistics which, however, is not of primary importance in an introductory course. Some of the elements of the subject are considered at suitable places in succeeding chapters, and a summary is made in the final chapter.

If statistical data could be selected by methods of random sampling as in laboratory experimentation, it would be much easier to apply theoretical measures of reliability and the "probable error," but, unfortunately, data gathered in field work cannot always be obtained entirely without bias. For example, a straw vote collected by a given publication designed to show the political strength of given parties may be biased by the fact that the publication goes chiefly to certain classes or occupational groups which are more or less favorable to one party as against another. If an attempt is made to avoid such bias by choosing, for example, names from a telephone directory, bias may again enter in that those who do not have telephones are not represented. In general, random observations are notoriously distorted by personal bias. People tend to observe and record facts favorable to their prejudices, and to discount or overlook facts that are unfavorable. Even supposedly scientific research has been rendered invalid by such unconscious bias. It may be taken for granted that there are no mathematical rules which can adequately measure such bias, and only by the most rigorous care can it be reduced to a minimum.

Primary and secondary data.—In the gathering of data a distinction may be made between primary and secondary sources. The use of primary sources may involve the arduous and often expensive methods of the personal interview and questionnaire, or perhaps the laborious transcription of data from original records in the books of business firms. It only too often happens that primary investigations prove ineffective because the figures collected are incomplete, or the questions asked are ambiguous or not comprehensive enough. Such failures can

be avoided only by thinking the whole problem through clearly in advance in terms of the final objective and the means available. The inductive investigative method calls for deductive abstract thought in its inception.

Investigations of the second type employ the data already collected and perhaps already analyzed in part or in whole. Such secondary sources may be census reports, or similar studies, the data of which are to be reworked for the purpose of drawing further conclusions such as may depend upon averages, trends, or correlations. In such studies the cautions regarding accuracy and comparability of data will be pertinent. It is very easy to draw unwarranted conclusions from incomplete or inaccurate data, or by means of invalid comparisons. For such purposes a thorough knowledge of the logic of statistical measurements and their limitations is indispensable.

Accuracy of calculation.—In order to achieve accuracy in calculation, it is very important that methods of checking be used. Additions and subtractions may be readily checked from the tape of the adding machine by proof-reading from the initial clearing sign to the final total, and multiplications and divisions should be checked by reading the numbers back to the copy after they have been placed upon the calculating machine. Figures copied from statistical tables, such as Barlow's "Tables of Squares, Cubes, etc.," and Hodgman's "Mathematical Tables," should also be carefully proof-read. It is generally possible to make a rough check of calculations by means of charts, as in the case of averages and trends, and in many cases the footings of columns in calculations will give a check.* Finally, when a solution is completed, it may be roughly checked by mental calculations and by the logic of the problem to make sure that it is plausible. Nevertheless, while a reasonable degree of accuracy is essential, time may be wasted in an attempt to secure spurious accuracy. Results obtained from a sample are not likely to be more accurate than the items of the

* Thus if deviations from an arithmetic average are taken, the algebraic sum is zero; if each item in a series is divided by a constant, then the total of the column, divided by the same constant, will check with the total of the quotients. A series of positive and negative deviations from the average, divided through by the average of the deviations, should give an absolute (i.e., negative signs disregarded) average of unity. A trend series should total practically the same sum as the series to which it is fitted. When one series is multiplied or divided by another series, however (i.e., the first item of one by the first item of the other, etc.), the averages of the series will not necessarily bear the same relation as the items, though they may approximate it. Many similar methods of checking will be observed by the student in particular cases.

sample itself. Hence calculations involving accuracy to several significant figures (figures of a "round" number excluding succeeding zeros and a decimal denominator) usually involve a waste of time. As a rule, the degree of accuracy may be determined by reference to the effect upon a chart designed to present the data and the results. A degree of accuracy beyond that which will register visibly on the chart is usually unnecessary. However, care must be taken in cases where errors are likely to accumulate through successive steps of calculation. For example, in trend fitting it is often necessary to add a constant successively, perhaps fifty or sixty times; or similarly to multiply by a constant successively a large number of times. In such cases the constant may be written to several more decimal places than will be used in the final figures in order to avoid a significant error at the end of the series. There will also be processes depending on the measure of a variable where perhaps the variability appears only in the fifth or sixth decimal place or significant figure. In such a case, accuracy must obviously be carried beyond this point. But for the most part, numbers rounded to three or four significant figures are sufficiently accurate for practical purposes.

In rounding numbers, the rule of business in rounding to the nearest cent is usually followed; that is, a remainder less than half a unit is disregarded, while half or more than half is counted as an additional unit. However, in order to avoid an exaggerated total where there are many half-units, this rule is sometimes revised to read: drop less than half, add one for a fraction over half, and change exactly half to the nearest *even* number. Thus 175,250 written as thousands is 175; 175,750 is 176; and 175,500 (half way between 175 and 176) is 176, while 174,500 is 174.

Tabulating the data.—Data are tabulated in various ways according to their nature. The order of arrangement of the classes into which the data are sorted may be alphabetical, chronological, geographical, by size or value, or by any other criterion that seems best suited to the case. In entering the data in the appropriate columns a certain amount of editing will often be necessary to reduce the figures to convenient units and to eliminate inconsistencies. The subdivision of the captions (headings) and stubs (left-margin designations) will allow many variations for which no definite rules can be laid down. The preparation of tables may best be studied by reference to census volumes and other compilations put out by reputable bureaus.

Frequency distributions.—From the standpoint of statistical analysis the most important form of tabulation is the so-called frequency distribution. This is obtained by sorting the data into classes according to

magnitude.* The classes should be so arranged that the data will tend to average near the middle of the class rather than to "bunch" near one extreme. To illustrate, suppose there is at hand a memorandum of the daily earnings of a hundred piece workers. The earnings range from \$2 to \$10, and are written to the nearest cent. They may be counted by classes, as in Table 1.

TABLE 1

A simplified frequency tabulation

Wage classes (Lower and upper limits)	Number of workers (Frequency)
\$2.00–\$3.99	20
4.00– 5.99	40
6.00– 7.99	30
8.00– 9.99	10

The numbers \$2, \$4, \$6, and \$8 are the lower limits (L_1) of the classes, and \$3.99, \$5.99, etc., are the upper limits (L_2). The number of workers in each class is the "frequency" (f). For purposes of precise calculation, the class limits are \$1.995 to \$3.995; \$3.995 to \$5.995; etc., splitting just between two possible magnitudes of the data. But in practice this is an unnecessary refinement when the space between two magnitudes is small, and generally the classes would be considered \$2–\$4; \$4–\$6; etc., understood to mean \$2 up to but not including \$4; etc. The mid-point of the class, that is, the average of the lower and upper limits, is usually called the class mark (m). For purposes of calculation the limits should always meet, so that the upper limit of one class is the lower limit of the next. The difference between the lower limit and the upper limit thus taken is called the class interval (i). The class interval should preferably be uniform throughout the table, though at times it may be necessary to disregard this rule.

* In statistical laboratories, government bureaus, and large business firms, the work of tabulating is generally done by machines. For this purpose, standardized cards are punched in such a way as to indicate the given data according to a prearranged code. The cards are then passed through a machine which sorts them into required classes and perhaps records or even calculates required results. In census bureaus very complex machinery is employed. The student may, however, learn the general principles of tabulation and the calculations based upon them from the study of statistical methods. The technique of operating machines is another matter, and can be learned effectively only by a course of training based upon the operation of the machine. The use of the more ordinary adding and calculating machines, such as are generally employed in statistical laboratories, may be learned under supervision or through the instruction manuals prepared by the manufacturers of the machines.

Plotting a frequency distribution.—A frequency tabulation may be plotted as a rectangular polygon, or histogram, with flat lines extended at the appropriate height above the class interval to represent the frequencies; or as a line polygon, where each frequency is plotted as a

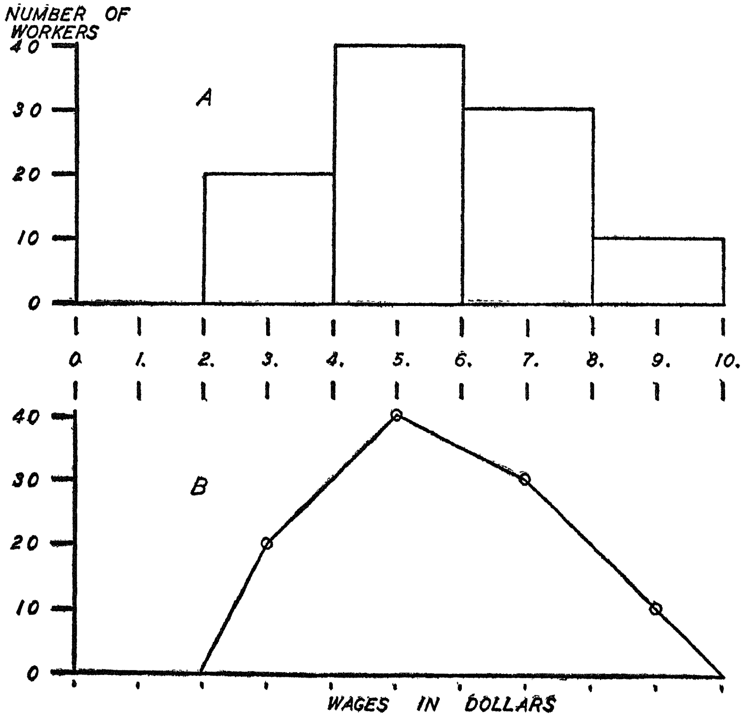


CHART 1

Graphic representation of the simplified frequency tabulation of Table 1. (*A*) Rectangular frequency chart presenting each frequency as a rectangle whose base is the specific class interval and whose height is the number of workers. This chart is often drawn as a mere outline omitting the portions of the rectangle which fall within the figure (cf. Chart 3). (*B*) Polygon frequency chart or line chart, formed by guide lines connecting points representing the frequencies plotted above their respective class marks. At the extremes of this figure, lines are drawn to the lowest and highest class limits respectively, thus representing accurately the range. Often, however, these two lines, if drawn at all, are drawn to the adjacent class marks where the frequencies are zero, although such an arrangement exaggerates the range, and may prove impracticable when it carries a tabulation into negative magnitudes. The rectangular form is generally preferable in that it represents correctly the area of each class within the assigned limits, but the polygon form may be more convenient if two or three charts are superimposed for purposes of comparison.

point above its class mark, and the points thus determined are connected by straight lines (cf. Chart 1). The former method is preferable except where two or more such distributions are plotted to the same scale. In plotting classes having a larger class interval than the pre-

dominant one, the height of the frequency should be reduced in the ratio that the interval is increased, in order that the area of the frequency may be correctly scaled.

For example, suppose that the frequencies of the classes 2 to 4, 4 to 6, 6 to 8, have been plotted in rectangular form, and the frequency 6 in the class 8 to 12, which has twice the regular interval, is to be plotted next. In order to give this frequency its proportional area, it should be plotted at one-half its given height, or 3, as if it were two classes of 3 each. In this way the frequencies are spread out proportionately over the larger class interval and the area of the frequency is commensurate with the areas of preceding frequencies. If the last class is an open class, that is, if it is given as 8 and above, the class interval is indeterminate. The interval, however, can be roughly estimated on the chart by extending it to such a degree that the adjusted height of the frequency will appear normal with reference to the descending size of the preceding frequencies. Or, on the same principle, it may be broken up into two or more classes of normal intervals, having successive frequencies of decreasing size totaling the given frequency of the open class. The relative size of these classes may be estimated by means of a table of the normal curve.

For purposes of calculation and more elaborate graphic representation, the frequency tabulation may be written to show the class limits (L_1 and L_2), the class marks (m), the frequencies (f) often expressed as percentages, the total frequencies (n), and the cumulative frequencies (Σ_1 and Σ_2), representing the total frequencies from the beginning of the

TABLE 2

A frequency tabulation arranged for computation. The cumulatives, Σ_1 and Σ_2, are the sub-totals of the frequencies, corresponding to L_1 and L_2.

Class, L_1 and L_2	Class mark, m	Frequency, f	Cumulatives, Σ_1 and Σ_2
\$2– \$4	\$3	20	0– 20
4– 6	5	40	20– 60
6– 8	7	30	60– 90
8– 10	9	10	90–100
		$n=100$	

table to L_1 and L_2, respectively, of a given class, as in Table 2. However, Σ_1 is often omitted, and Σ_2 appears as Σf, or merely Σ (cf. Table 3, p. 16).

Plotting a cumulative curve.—Methods of graphing the frequencies of Table 2 have already been suggested in Chart 1, and methods of graphing the cumulatives (Σ_1 and Σ_2) may now be considered. In Chart 2 a method of graphing the cumulatives (lower figure, *B*) is

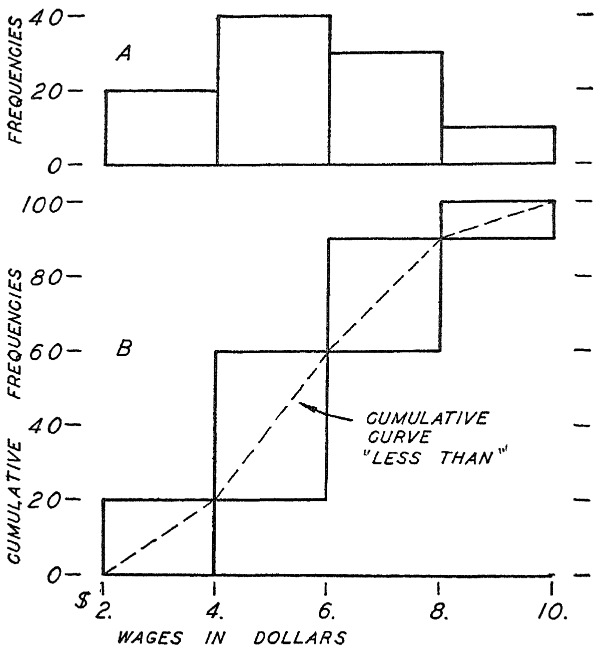


CHART 2

A comparison of frequency and summation charts: graphic representation of the data of Table 2. (*A*) Rectangular frequency chart, similar to Chart 1 (*A*), the spread of each class being represented on the horizontal scale and the number of frequencies on the vertical scale. (*B*) Cumulative chart showing the spread of each class on the horizontal scale and the spread of the cumulatives on the vertical scale. The diagonals connecting the rectangles form the so-called cumulative curve, which is commonly drawn without the rectangular frequencies enclosing it. It is sometimes designated as the "less than" curve to distinguish it from the "more than" curve which may be similarly drawn by cumulating the frequencies beginning at the other end of the distribution. Any point on the curve may be taken to indicate approximately the number of workers, as indicated on the vertical scale, receiving "less than" the wage designated on the horizontal scale below.

depicted as related to a rectangular chart of the frequencies (upper figure, *A*). It will be seen that the distribution is first represented as before by rectangles expressing the class limits by their width (horizontal scale) and the frequencies by their height (vertical scale). The cumulative chart is constructed, in effect, by lifting in succession each rectangle except the first to a position where the lower left-hand corner coincides

with the upper right-hand corner of the preceding rectangle. Diagonals are then drawn through the successive rectangles. In actual charting of the cumulatives, however, the rectangles are usually omitted and the diagonals drawn as a broken curve. This may readily be done directly by plotting the first and second summations (Σ_1 and Σ_2 on vertical scale) against the lower and upper class limits (L_1 and L_2 on horizontal scale). Or, since the series thus designated include many duplicates, the directions for plotting may be described more simply as a plotting of Σ_2 against L_2, beginning with a zero summation at the first L_1. The broken curve thus obtained—identical with the diagonals of Chart 2—is ordinarily called a "less than" cumulative (or ogive). The designation "less than" refers to the reading of any point on the curve. On the assumption that the distribution within each class is regular, as pictured by the rectangular distribution, any point on the curve read on the coordinate scales indicates the number of workers (reading on the vertical scale) receiving wages less than the indicated wage (reading on the horizontal scale). This characteristic of the cumulative curve is made use of later in certain calculations. In a similar way a "more than" summation curve may be constructed by cumulating the frequencies in reverse order from the end of the distribution rather than from the beginning. Graphically, this would be equivalent to raising the rectangles in the frequency distribution in the same manner but in the reverse order from that indicated in the chart. The former method is the more convenient, however, and is the one most commonly adopted.

Summation curve and bar chart.—The cumulative chart in the form here described is not of great value for popular presentation of frequency distributions, but is chiefly useful in connection with certain calculations, as just suggested. It may, however, easily be transposed into a very simple form of chart by reference either to the original data or to a simplified and smoothed form of these data. This type of graph is illustrated in Chart 2*a*. For the purpose of constructing this chart, let us assume that the wage distribution of Table 2, as plotted in Chart 2, has been reduced proportionately to the following items: first worker, \$2.50; second worker, \$3.50; third worker, \$4.25; fourth worker, \$4.75; fifth worker, \$5.25; sixth worker, \$5.75; seventh worker, \$6.33; eighth worker, \$7.00; ninth worker, \$7.67; tenth worker, \$9.00. An inspection of these figures will show that they fall into the classes indicated in Tables 1 and 2 and that they represent the same frequencies reduced one-tenth. Within each class they are regularly spaced so that there is an equal interval between the items, except adjacent to the class limits where the spacing is one-half of the interval used in the given

class. If, now, these ten items are plotted successively as bars of uniform width, with the lengths adjusted to the horizontal wage scale in dollars, there is obtained a simple chart of the individual workers, or of representative workers graded from the highest paid to the lowest paid. If this chart is compared with the cumulative curve of Chart 2, it will be

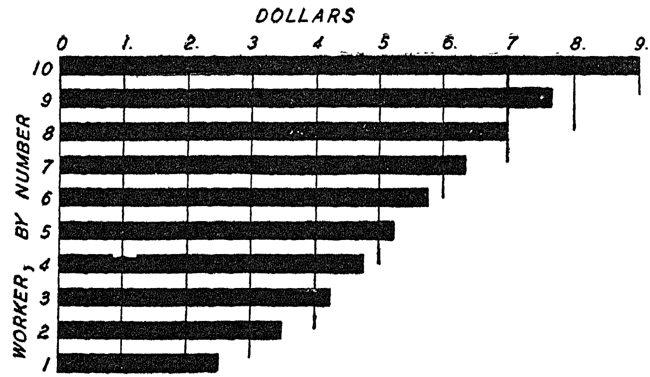


CHART 2*a*

Bar chart of a frequency distribution of ten workers in the tabulation $m = 3, 5, 7, 9$, and $f = 2, 4, 3, 1$, which is the same as Table 2 and Chart 2 (pp. 11 and 12) except that, for convenience, the number of workers has been reduced to ten. The wages charted are as follows: \$2.50; \$3.50; \$4.25; \$4.75; \$5.25; \$5.75; \$6.33; \$7.00; \$7.67; \$9.00. If the ends of each bar are connected, a cumulative curve is drawn similar to that of the original distribution, but indicating an indefinitely large number of workers.

seen that it is closely analogous to it, in that the ends of the bars trace a cumulative curve, and therefore could have been drawn from a chart of that curve with any desired number of workers taken as representative. Hence, a cumulative curve may be interpreted as simply an outline of a smoothed bar chart, in which case it obviously may be assumed to represent a very large number of frequencies distributed smoothly according to the general scheme indicated by the tabulated distribution. When the curve is thus interpreted, it becomes much more intelligible to the inexpert reader as a picture of the distribution. Both the cumulative curve and the bar chart based upon it may be further smoothed, if desired, by using an irregular curve and drawing a smoothed line between the plotted points, in such a way that the angles formed at the class limits will be eliminated. A somewhat similar smoothing may be applied to a frequency distribution, as is indicated in Chart 3, and more accurate methods of arriving at it will be considered later.

Types of distributions. (*a*) *The normal frequency curve.*—Frequency tabulations commonly follow one of two types, the first and most impor-

tant of which is the so-called normal frequency distribution, including the simpler binomial distribution from which, in a certain sense, it is derived. The binomial distribution may be introduced by reference to the elementary data of Table 3, which is assumed to represent a

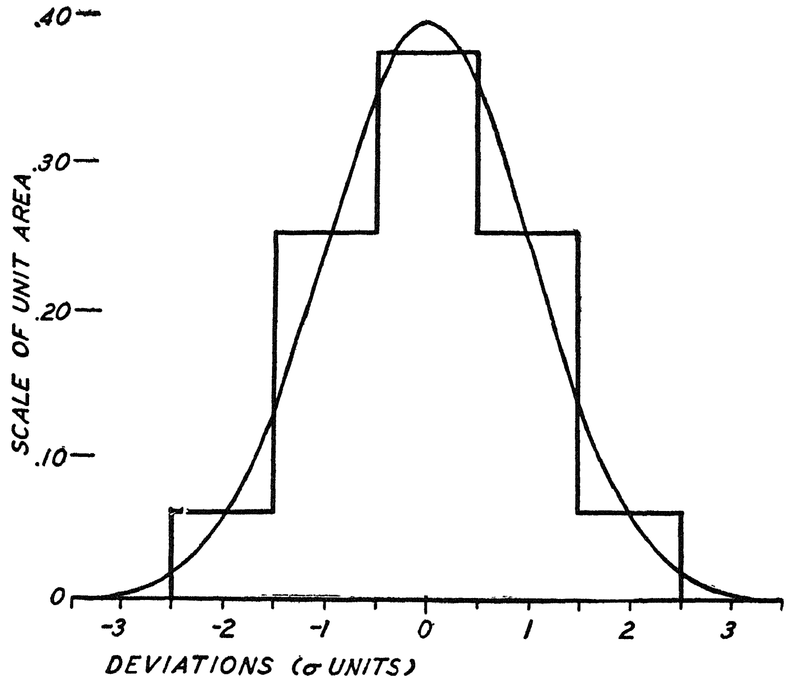


CHART 3

Graphic representation of a binomial distribution of five classes (rectangular graph) expressed in percentages of a total area, the base line being expressed in units of one class interval each, which, in this case, is also the standard deviation unit, to be explained later. The smooth line presents the curve as it would appear with an infinite number of classes, that is, the normal curve of distribution, or normal probability curve.

tabulation of workers in a given factory according to daily wages received. The table shows a classification of the workers (frequencies) characterized by greater regularity than would ordinarily be found in practice. The ratios of the frequencies, in hundreds, 1 : 4 : 6 : 4 : 1, are in fact an exact theoretical distribution of five discrete classes as computed by the mathematician according to the laws of chance, as when four coins are thrown simultaneously sixteen times and the number of heads turning up at each throw are counted and tabulated. Such an experiment will approximate, as a rule, the following distribution: no

heads, once; one head, four times; two heads, six times; three heads, four times; and four heads, once. This theoretical expectation follows

TABLE 3

The binomial frequency distribution. Assumed distribution of daily wages in factory X, May, 1930

(1) Class limits, L_1 and L_2	(2) Mid-class wage (m)	(3) Number of workers (f)	(4) Sub-total (Σf)
\$ 2 to \$ 4	\$ 3	100	100
4 to 6	5	400	500
6 to 8	7	600	1100
8 to 10	9	400	1500
10 to 12	11	100	1600
		1600	

the ratios of the coefficients of the terms of a binomial, such as $a + b$, raised to the fourth power. For other distributions (three to nine classes) the ratios of the theoretical frequencies are given in Table 4.

TABLE 4

Binomial distributions of three to nine classes. Each distribution after the first is a summation of the preceding, with a unit term appended. The number of classes is one more than the power of the binomial (s); that is, $s + 1$. The total frequencies may be found as $\Sigma f = n = 2^s$.

Binomial and power (s)	Number of classes ($s+1$)	Ratio of frequencies	Total frequencies (n)
$(a+b)^2$	3	1 : 2 : 1	4
$(a+b)^3$	4	1 : 3 : 3 : 1	8
$(a+b)^4$	5	1 : 4 : 6 : 4 : 1	16
$(a+b)^5$	6	1 : 5 : 10 : 10 : 5 : 1	32
$(a+b)^6$	7	1 : 6 : 15 : 20 : 15 : 6 : 1	64
$(a+b)^7$	8	1 : 7 : 21 : 35 : 35 : 21 : 7 : 1	128
$(a+b)^8$	9	1 : 8 : 28 : 56 : 70 : 56 : 28 : 8 : 1	256

It will be seen that in Table 4 the frequencies increase symmetrically from each extreme toward the maximum central frequency, or mode. As the number of classes is increased, the curve as plotted in a rect-

angular form shows relatively smaller graduated steps from one class to the next, and a smoothed line called the normal curve is approached more and more closely. The normal curve may therefore be regarded theoretically as a binomial distribution, or point binomial as it is sometimes called, in which the number of classes has been increased infinitely to make the smoothed curve.* In the smoothed curve the class interval of a theoretical five-class distribution is taken as a unit of spread, measuring from the central class mark, as is pictured in Chart 3.

(*b*) *The logarithmic frequency distribution.*—In practice, the frequencies of a tabulation are commonly found to be distributed in a series that extends farther from the largest frequency (mode) on one side than on the other. Such a distribution is said to be skewed, and if it approximates the form of a normal distribution when plotted on a horizontal logarithmic scale, it is said to be a logarithmic normal distribution. Such a distribution is illustrated in Table 5, and is plotted in Chart 4. In this case the distribution is said to be positively skewed; that is, the frequencies extend further from the mode into the upper magnitudes than into the lower. In Chart 5 the same distribution is plotted on a horizontal logarithmic scale, and a smooth curve is fitted to the frequencies by methods explained in Chapter IX. As the chart indicates, the curve thus fitted takes the form of a normal distribution when plotted on the horizontal logarithmic scale. This tendency for a skewed dis-

* The binomial distribution, or point binomial, for $s + 1$ classes, where n is the total frequencies, is expressed as,

$$n(\tfrac{1}{2} + \tfrac{1}{2})^s$$

or, for skewed distributions, as

$$n(p + q)^s$$

where $p + q = 1$. The normal curve of distribution of area n, may be expressed mathematically by the equation

$$Y = \frac{n}{\sqrt{2\pi}} e^{-x^2/2}$$

Where e is the mathematical constant 2.718 (cf. Appendix), and x is expressed in units of the class intervals of a five-class distribution measured from the center. Mathematically, the normal curve is the limit which is approached by the point binomial as the number of classes is increased. If a given number of classes are artificially made out of continuous data conforming to the smoothed curve, the frequencies will not necessarily conform exactly to the binomial classification. There are many other types of distribution (cf. Rietz, "Mathematical Statistics," Chapters I–III), but experience shows that the data of the social sciences may generally be described well enough for most purposes by some adaptation of the normal or the logarithmic normal curves, including the binomial as an approximation to the normal. Methods of fitting these curves, together with the theories of probability which they represent, are considered in Chapter IX.

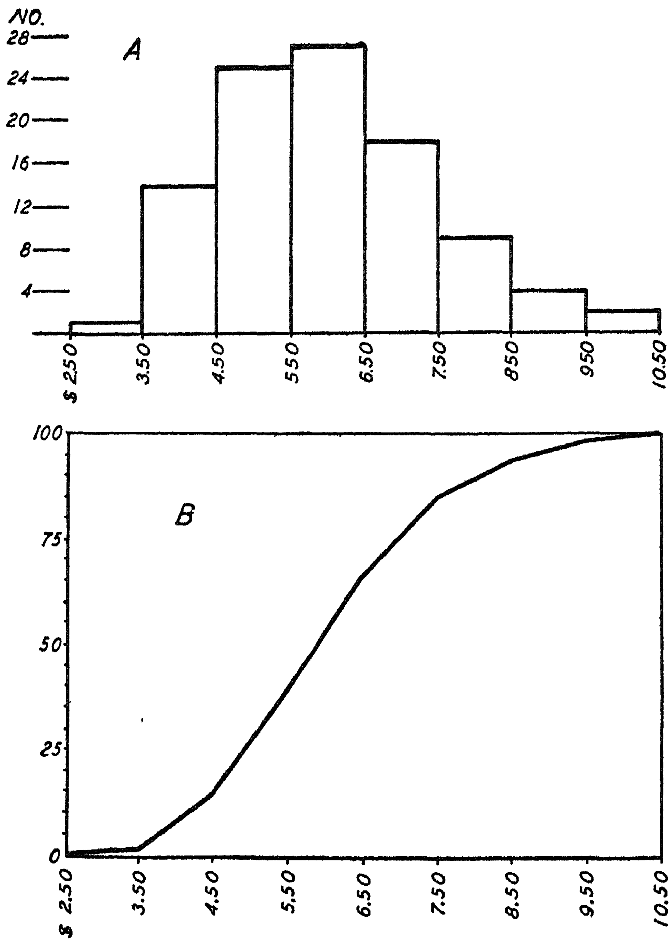


CHART 4

(A) Graphic representation of the data of Table 5, each frequency (%) being indicated by a rectangle extending horizontally from the lower to the upper class limit and vertically to the height indicated by the frequency. (B) Graphic representation of the cumulative percentage frequencies of Table 5, each cumulative being plotted against its appropriate class limit. That is, each Σ_2 (vertical scale) is plotted against its corresponding L_2 (horizontal scale), beginning with zero against the first L_1. The summation curve is an integration of the distribution curve as plotted.

tribution to approach the normal in form when thus plotted is a crude test of the logarithmic type.

TABLE 5

The logarithmic type of frequency tabulation. Distribution of daily earnings in factory Y, May, 1930. In later calculations based on this table the percentage frequencies are substituted for the actual. For some purposes the logarithms of the class limits and the class marks are required, as given below.

Class limits, L_1 and L_2	Class mark, m	Frequency, f	Percentage frequency, f (%)	Cumulative frequencies Σf	Cumulative frequencies Σf (%)
\$2.50–\$ 3.50	\$ 3.00	5	1	5	1
3.50– 4.50	4 00	70	14	75	15
4.50– 5.50	5.00	125	25	200	40
5.50– 6.50	6.00	135	27	335	67
6.50– 7.50	7.00	90	18	425	85
7.50– 8.50	8 00	45	9	470	94
8.50– 9 50	9.00	20	4	490	98
9.50– 10.50	10.00	10	2	500	100
		$n = 500$	100		

L and m	Log	L and m	Log	L and m	Log
2 5	0.3979	5.5	0.7404	8.5	0.9294
3.0	.4771	6.0	.7782	9.0	0.9542
3.5	.5441	6.5	.8129	9.5	0 9777
4.0	.6021	7.0	.8451	10.0	1.0000
4.5	.6532	7.5	.8751	10.5	1.0212
5.0	.6990	8.0	.9031		

Double frequency tabulation.—It is sometimes necessary to tabulate data with respect to two or more different distributions at the same time. Any two of these distributions, whether logarithmic, normal, or irregular, may then be compared by means of a double frequency tabulation. For example, suppose that certain cities (cf. Table 6) have been studied with respect to their ratings for certain vital, social, or economic phenomena, and percentages measuring their status according to the two required characteristics have been obtained (cf. Table 6, Index A and Index B). A scale of classification, with class limits and class marks, is arranged for each index separately; in the accompanying illustration Index A is distributed in the classes 3.5–7.5; 7.5–11.5; 11.5–15.5; 15.5–19.5; 19.5–23.5, of which the class marks are 5.5;

9.5; 13.5; 17.5; and 21.5. In a similar way Index B is distributed in the class intervals 2.5–7.5; 7.5–12.5; 12.5–17.5; 17.5–22.5; 22.5–27.5, of which the class marks are 5, 10, 15, 20, and 25. The double frequency tabulation may be made as indicated, the tallies being entered in the appropriate cells of the table. The cells picture the distribution accord-

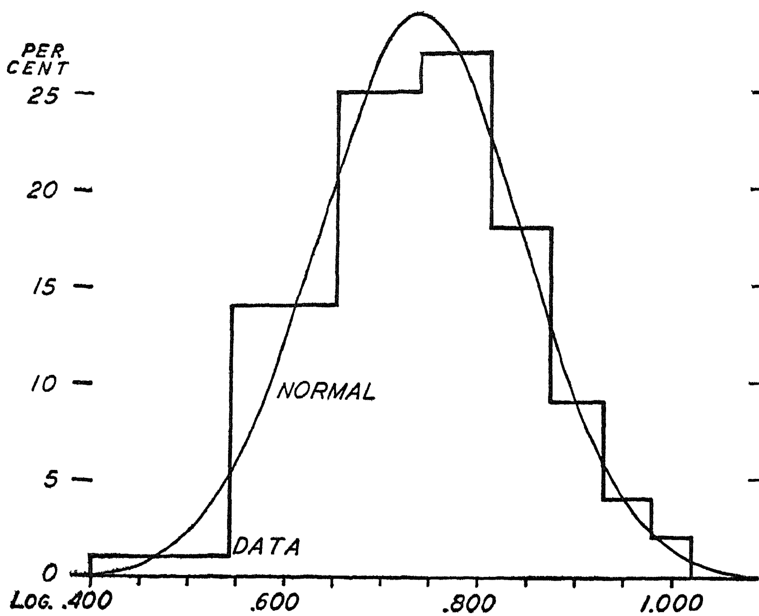


CHART 5

Graphic representation of the data of Table 5 plotted to a logarithmic horizontal scale. This is done by replacing the class limits and class marks by their logarithms, and plotting these logarithms on the horizontal scale. As a result the width of the rectangles is progressively lessened. A true logarithmic normal curve thus plotted may be smoothed to the normal form as indicated by the smoothed line, which is similar in form to the smoothed line of Chart 3. The method of calculating the smoothed line is explained in Chapter IX. When plotted on the ordinary X-scale, this smoothed line is skewed to the right, together with the data.

ing to both scales, the Y-scale being written with the highest values at the top of the column in order to make it correspond to the usual graphic representation. The Y-frequencies are totaled by rows to the right, and the X-frequencies are totaled by columns at the bottom of the table. The two tabulations may, of course, be written separately for purposes of further analysis.

Tabulating time series.—When the data consist of time series, such as index numbers of wholesale prices by months through a period of years, tabulation usually consists merely of presenting the series in any

TABLE 6

A double frequency tabulation: Cities A to P rated for certain social characteristics as indicated by Index A and Index B. I. Original tabulation of data. II. Data tabulated in five-class distributions, the separate sets of frequencies appearing as row and column totals. III. The tabulations arranged separately for purposes of further analysis.

I. *Original data.*

	Index A	Index B		Index A	Index B
A	13	19	*I*	22	24
B	17	16	*J*	14	16
C	13	15	*K*	9	14
D	19	25	*L*	9	11
E	14	8	*M*	14	14
E	10	6	*N*	6	4
G	11	12	*O*	19	21
H	19	20	*P*	15	15

II. *Double frequency tabulation.*

Index B		Index A: Class and m					
Class	m	3.5–7.5	7.5–11.5	11 5–15 5	15.5–19 5	19.5–23.5	
		5.5	9.5	13 5	17.5	21.5	
22.5–27.5	25				/	/	2
17.5–22,5	20			/	/ /		3
12.5–17.5	15		/	/ / / /	/		6
7.5–12.5	10		/ /	/			3
2.5– 7.5	5	/	/				2
		1	4	6	4	1	16

III. *Separate tabulations.*

Index A			Index B		
Class	m	f	Class	m	f
3.5– 7.5	5.5	1	2.5– 7.5	5	2
7.5–11.5	9.5	4	7.5–12.5	10	3
11.5–15.5	13.5	6	12.5–17.5	15	6
15 5–19.5	17.5	4	17.5–22.5	20	3
19.5–23.5	21.5	1	22.5–27.5	25	2

consecutive arrangement suitable to the case at hand. If the number of years is large, they may be written consecutively down the "stub" (left-hand margin), and the months may be similarly written across the top. If the number of years is small the two scales may be interchanged. When such tables are presented for the purpose of emphasizing current data, they are often arranged with the scales reversed so that the latest item appears in the upper left-hand portion of the table. Examples of such tabulation may be found in the business services and in the *Survey of Current Business.* Occasionally items in a time series will be compiled in a frequency tabulation, but this will usually follow complex calculations like trend fitting.

The plotting of time series usually takes a form analogous to the polygon frequency distribution or line chart indicated in Chart 1(B), p. 10, though occasionally the rectangular form or separate bars may be used. In such charts the vertical scale is often logarithmic, so that in effect the logarithms of the data rather than the data themselves (vertical scale) are plotted against the actual years (horizontal scale). Such charting, however, may be more conveniently done on common semi-logarithmic or ratio paper, in which a numerical vertical scale, logarithmically spaced, is employed. In using such paper the scale appearing in the margin may be multiplied by any convenient constant, but it should not be changed by an addend, since an addend vitiates the ratio spacing. It should be observed that just as the ratio scale in Chart 5 distorts the width of the frequencies, so, in the same way, the logarithmic scale distorts the height of the plotted time series. There is an advantage, however, in this distortion since it results in giving equal slope to equal ratio changes, as is illustrated in Chart 6 by the compound interest line (I_1) plotted on an arithmetic scale, and the same line (I_2) plotted on a ratio scale. On a ratio chart, also, percentage fluctuations, such as the ups and downs of business in a large company and in a small company, may be directly compared as to the relative amount of change. Such a comparison is illustrated by the lines A_2 and B_2 in Chart 6, which are assumed to represent production in a large business and a small one, respectively. The fluctuations, or cycles, plotted to a ratio scale, appear graphically to be the same in degree—an appearance which is confirmed by reference to the data; that is, each rise is a 50% rise. When, however, the same cycles are plotted to an arithmetic scale (lines A_1 and B_1) they appear much larger in the larger business, as in fact they are in absolute amounts.

An inspection of charts appearing in the business services and in current periodicals employing statistical data will show that the ratio scale is very commonly employed. It may readily be distinguished

by the progressive narrowing of the spaces representing equal numerical increments upward on the chart. Ratio charts are of course not adapted to scales which may drop into negative numbers, since there

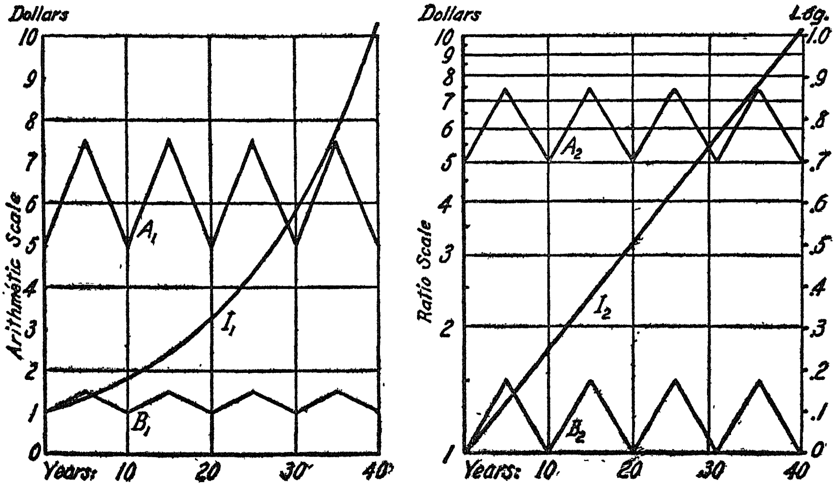


CHART 6

Time series plotted on arithmetic and ratio scales (cf. data below). It will be seen that the cycles *A* and *B* change in the same ratio and therefore appear similar in the ratio chart. The compound interest series, *I* (the amount of one dollar compounded annually at 6% for 40 years), appears as a straight line on the ratio chart because the rate of change is constant. The left-hand scale on the ratio chart is similar to that commonly printed on ratio paper, but the logarithmic scale from which it is formed is presented for purposes of comparison on the right-hand margin. The ratio scale is measured by ruling the line 2 at the log of 2 (0.3010), the number 3 at the log of 3 (0.4771), etc. It will be noted that on the ratio scale a geometric series, such as 1, 2, 4, 8, etc., or 1, 3, 9, etc., is spaced at equal intervals. For convenience, the ratio scale may be multiplied by any suitable constant.

Years from origin	Cycle *A*	Cycle *B*	Amount (*I*)	Log *I*
0	$5.00	$1.00	1.000	0.0000
5	7.50	1.50	1.338	0.1265
10	5.00	1.00	1.791	0.2531
15	7.50	1.50	2.397	0.3796
20	5.00	1.00	3.207	0.5061
25	7.50	1.50	4.292	0.6326
30	5.00	1.00	5.743	0.7592
35	7.50	1.50	7.686	0.8857
40	5.00	1.00	10.286	1.0122

can be no zero on a ratio scale. In general, their use should be limited to those cases where ratio changes are to be emphasized or contrasted.

Graphic representation.—A brief discussion has already been given of certain elementary charts such as are adapted to frequency distribu-

tions and time series. No attempt will be made at this point to give a detailed exposition of methods of graphic presentation, since the subject is better studied in connection with specific problems as they occur, as in succeeding pages where various charts covering the most common graphic forms will be found. Further study may best be done by reference to the excellent volumes on the subject now available to the student (*e.g.*, Karsten: "Charts and Graphs," and similar specialized works). A few general rules or precautions may, however, be emphasized.

In the first place it should be remembered that the vertical and horizontal scales are not usually in comparable units, and therefore no mathematical rule governs their relative spread. Hence the arrangement of the two scales should be such as to give the chart reasonably good proportions, the typical proportions being a height of perhaps two-thirds the width. This is by no means a rigid rule, however, and may be changed with varying problems; but in general the proportions of a 3 by 5 index card may be taken as a desirable norm. Certainly the chart should not be distorted in order to create an exaggerated impression.

One way in which exaggeration on arithmetic scales may sometimes be avoided is by drawing the chart so that the zero line appears as the base line. If index numbers are plotted, the 100% line may also be drawn heavier in order to emphasize the space from zero to one hundred. If this is done, the actual change recorded by the fluctuations of the data will be seen in terms of their relation to the zero base line; hence the proportional change will be visibly suggested. In frequency distributions scattering close to zero as a class limit, the same rule may also be applied to the vertical axis at the left margin, so as to give a more accurate impression of the relative spread. However, it often happens that the nature of the data does not permit a representation of the axes. And again, the use of negative numbers places the zero coordinate lines within the chart rather than at the margin. About all that can be said, then, is that only important lines of reference should be stressed, and that the marginal scale lines should coincide with the axes when this is feasible; otherwise, they should not be stressed. In the latter case the scale may be indicated merely by dashes, or a break may be pictured to suggest the fact that the vertical scale is incomplete. When the scale is represented merely by dashes without guide lines drawn through the chart, it is usually desirable to draw dashes corresponding to the scale at the upper and right-hand margins so that the chart can be read by means of a ruler.

In charts where it seems necessary to plot two separate sets of data—as wages in a certain industry compared with the productivity of that

industry through a series of years—separate scales may be arranged for each series such as to allow the appropriate juxtaposition of the two lines. Sometimes both of the scales are placed in the left-hand margin; sometimes they are placed at the opposite margins. Occasionally three or four sets of data may be plotted in a given chart, but this is likely to make the chart unduly complex.

In all graphic presentation care should be taken to indicate clearly the units plotted either by an accompanying "legend," or by a label, to which may be affixed an arrow pointing to a specified line. When several series of data are plotted in the same chart, they should be clearly distinguished by solid, broken, and dotted lines, or other devices. It is desirable also that the data and the sources from which they are obtained appear on or in connection with the chart. Clear and concise titles and descriptive labels should also be attached.

In popular literature, graphs are often presented in pictorial form; that is, instead of bars or lines, drawings of ships, men, bales of cotton, or other units are presented. There is great danger of ambiguity, however, when such a method is employed. If, for example, the size of two navies is presented graphically by two battle ships, presumably proportioned to the size of the respective navies, the reader may assume that the ratio of the lengths of the two ships indicates the comparison of the navies. But if he assumes that the areas depicted represent the required ratio, then the ratio of the lengths is squared; while if he assumed that the volumes depicted are the measure, the ratio is cubed. Thus, whatever scheme of proportions may be adopted, ambiguity is almost certain to result. Pictorial graphic presentation is therefore seldom desirable except from the narrow point of view of partisan advertising or propaganda. If it is used for scientific purposes, the basis of the comparison should be stated. However, a less objectionable form is one in which the unit pictured is repeated in a row to form, in effect, a bar chart. In this case the units may be all alike; and the length of the rows correctly represents the relationship of the data. German and Austrian statistical bureaus have made a great deal of use of this type of chart in recent years. An example of such a pictorial chart appears later in a discussion of index numbers (cf. Chart 14*b*, p. 100).

The student of statistics should observe the graphic presentation appearing in authoritative current sources such as the "Statistical Atlas of the United States," published by the Bureau of the Census, or reputable business services and financial and social periodicals. In graphics, as in other fields, practice tends to change, and the student who observes the trend of the best practice is likely to be better informed than the one who blindly follows certain set rules.

EXERCISES

The accompanying exercises, together with those appended to later chapters, are chiefly simple problems based on inadequate data, and are presented in order to furnish practice sufficient to fix the method in mind. They should be supplemented by more extensive data chosen from *The Survey of Current Business*, *The Statistical Abstract*, and other available sources.

1. Tabulate the data on earnings given below, in earning classes of \$6–\$8 etc., under the headings given at the foot of the table.

Individual earnings per week in workshop X, 1925 and 1930, classified by occupations

Occupation	1925		1930	
	Workers	Earnings	Workers	Earnings
A	1	\$ 7.57	1	\$ 9.85
B	1	9.47	2	11.99
C	2	12.10	1	14.38
D	3	10.30	3	12.80
E	1	14.30	1	16.46
F	1	13.83	2	15.05
G	2	11.80	1	13.67
H	1	6.43	1	8.15
I	1	13.97	1	15.54
J	1	11.55	2	12.95
K	2	8.70	2	10.00
L	1	16.99	1	19.03
M	2	9.04	1	11.02
N	1	15.71	1	17.52

Tabulate under headings:

Class limits; Class mark; Frequency; Total wages (mf); Cumulative frequencies at L_2; Per cent frequencies. For certain graphic purposes the following columns are sometimes added: Per cent wages; Cumulative per cent wages.

2. (a) Business cycles in the United States arranged in chronological order (1796–1923) have had the following duration as measured to the nearest year (cf. Mitchell, "Business Cycles," National Bureau of Economic Research, p. 387): 6, 6, 5, 3, 7, 3, 3, 5, 4, 3, 6, 1, 2, 6, 4, 3, 5, 5, 4, 9, 5, 3, 2, 3, 4, 3, 4, 2, 3, 5, 2, 3. Tabulate in classes of one year each, and plot the distributions in the rectangular and polygon forms. Reduce to percentages, and write and plot the cumulatives.

(b) The corresponding figures for English business cycles are given below (1793–1920). Tabulate and plot as above; also combine the American and English data into one tabulation and plot: 4, 6, 4, 3, 5, 4, 6, 4, 2, 6, 10, 7, 4, 8, 8, 9, 8, 10, 7, 6, 5, 2.

3. (a) The following series is assumed to represent the daily earnings in dollars of 100 workers. Tabulate them in classes having the intervals \$2.50–\$3.50; \$3.50–\$4.50, etc., and plot in the usual rectangular and polygon forms. Write and plot the cumulatives.

10 25	7.03	8.00	4.25	9 75	4.76	7.78	7.89	4.32	4.88
5.36	5.00	4.80	5.93	3.61	4.11	4.56	6.30	5.12	6 41
4.60	3.54	5 85	6.26	4.68	5.74	3.82	3.89	7.14	7.47
6.64	4.46	6.37	9.12	3.75	6.19	5.40	4.64	5 81	5.48
4.92	6.04	4.52	9.38	5.44	7.25	5.96	5.04	5.89	7.42
5.08	7.31	3.00	6.92	5 78	8.11	6.15	5.59	6.75	5.70
5.67	8.44	5.52	4.84	6.00	8.33	6.58	6.07	6.86	5.20
7.19	5.24	6.53	3.68	7.36	6.44	8.22	6.33	5.63	4.72
5.16	3.96	7.56	6.11	5.56	7.67	8.88	6.69	5.32	4.18
4.04	6.81	8.62	7.08	4.96	5 28	6.97	6.22	4 39	6.48

(b) Tabulate the following items in classes of 5–7; 7–9; etc., and plot. Write and plot the cumulatives.

12.32	9.21	8.75	17.71	14.94	9.62	9.38	16.90	11.20	9.46
7.42	7.75	10.62	10.38	10.12	7.58	17.43	12.16	16.30	15.70
15.50	16.50	7.92	8 08	7.08	15 90	12.64	8.58	10.71	14.00
13.18	5.50	13.53	6.50	18.29	14.71	19.33	11.84	12.88	10.79
16.10	10.04	11.04	7.25	12.72	13.88	12.40	14.24	10.54	12.00
12.56	14.12	9.12	13.76	9.79	11.76	11 36	15 30	12.48	8.42
9.04	10.46	9.88	13 65	12 96	10.29	9.71	18 57	14.59	12.80
17.14	11.12	13.06	18.00	18 86	10.96	8 25	14 35	10.88	20.67
20.00	11.60	11.28	9.29	14.47	16 70	8.92	11.52	11.68	12.24
10.21	14.82	9.96	11.44	13.29	12.08	11 92	9.54	13.41	15.10

4. The series of items given below is taken to represent relatives (percentages) of wholesale prices of 20 important commodities in 1923 as compared with prices in 1913, which is taken as the base year.

(a) Tabulate the data in classes of 90–110; 110–130; etc., and plot in the rectangular form. Also write and plot the cumulatives.

(b) In the same way tabulate in classes of 45–55, etc., and plot the reciprocals of these percentages, also expressed as percentages; e.g., $1/142\% = 70.4\%$.

Series: 123, 166, 139, 133, 100, 156, 200, 145, 149, 115, 135, 158, 122, 188, 178, 128, 125, 147, 106, 169.

5. The following tables give a classification of wholesale price relatives (percentages) of individual commodities, on a 1913 base (Bureau of Labor Statistics data), for the years 1923, 1924, and 1925. The table indicates the number of relatives in each commodity group falling between the class limits indicated (30% and under 50%, etc.). The commodity groups are as follows: I, Farm products; II, Foods; III, Cloth and clothing; IV, Fuel and lighting; V, Metal and metal products; VI, Building materials; VII, Chemicals and drugs; VIII, House furnishings; IX, Miscellaneous.

(a) Plot as rectangular frequency charts the data for each year by groups and totals.

(b) Reduce the frequencies to percentages, and write and plot the cumulatives.

(*A*) Distribution of wholesale price relatives in 1923, on a 1913 base

Class	Commodity Group									
	I	II	III	IV	V	VI	VII	VIII	IX	Total
30– 50%	......	......	......	......	......	......	......	......	1	1
50– 70	......	1	......	......	......	......	1	......	0	2
70– 90	3	3	......	......	......	......	5	......	0	11
90–110	6	8	......	......	6	......	5	......	3	28
110–130	11	23	1	2	2	3	11	2	4	59
130–150	15	15	6	0	2	1	1	0	7	47
150–170	5	17	6	2	17	5	3	4	5	64
170–190	5	8	9	3	7	9	4	5	3	53
190–210	4	5	15	3	2	5	3	3	1	41
210–230	4	0	16	3	0	7	2	6	0	38
230–250	2	1	7	2	0	4	2	6	0	24
250–270	......	......	2	......	1	1	0	1	0	5
270–290	......	......	0	......	......	1	0	1	1	3
290–310	......	......	1	......	......	......	0	2	......	3
310–330	......	......	......	......	......	......	0	1	......	1
330–350	......	......	......	......	......	......	2	......	......	2
Total..	55	81	63	15	37	36	39	31	25	382

(*B*) Distribution of wholesale price relatives in 1924, on a 1913 base

Class	Commodity Group									
	I	II	III	IV	V	VI	VII	VIII	IX	Total
10– 30%	......	......	......	......	......	......	......	......	1	1
30– 50	......	......	......	......	......	......	......	......	0	0
50– 70	1	1	......	......	......	......	2	......	0	4
70– 90	2	2	......	......	1	......	5	......	0	10
90–110	4	6	......	1	3	......	9	......	5	28
110–130	6	19	1	0	4	3	6	2	4	45
130–150	17	22	7	3	12	3	2	2	7	75
150–170	13	16	9	3	10	4	4	5	3	67
170–190	5	9	12	3	5	12	2	6	3	57
190–210	2	5	15	0	1	7	4	4	0	38
210–230	3	0	15	3	0	6	2	4	1	34
230–250	1	1	3	0	0	2	0	2	0	9
250–270	1	......	2	1	1	1	2	2	1	11
270–290	0	......	0	1	......	......	0	1	......	2
290–310	0	......	1	......	......	......	0	3	......	4
310–330	0	......	......	......	......	......	0	......	......	0
330–350	0	......	......	......	......	......	0	......	......	0
350–370	0	......	......	......	......	......	0	......	......	0
370–390	0	......	......	......	......	......	0	......	......	0
390–410	0	......	......	......	......	......	1	......	......	1
410–430	0	......	......	......	......	......	......	......	......	0
430–450	1	......	......	......	......	......	......	......	......	1
Total..	56	81	65	15	37	38	39	31	25	387

(*C*) Distribution of wholesale price relatives in 1925, on a 1913 base

Class	Commodity Group									
	I	II	III	IV	V	VI	VII	VIII	IX	Total
50– 70%	......	1	......	......	......	......	2	......	......	3
70– 90	3	0	......	......	1	......	5	......	2	11
90–110	2	3	......	......	2	1	12	1	3	24
110–130	5	17	1	3	6	0	7	1	4	44
130–150	11	15	6	1	14	5	3	3	6	64
150–170	13	14	12	4	7	7	3	6	4	70
170–190	14	16	17	2	2	10	2	7	5	75
190–210	4	9	11	0	4	7	6	2	1	44
210–230	3	5	11	3	1	3	2	4	0	32
230–250	4	1	7	1	......	4	0	4	1	22
250–270	0	......	1	0	......	1	1	2	0	5
270–290	0	......	0	0	......	......	0	1	1	2
290–310	0	......	1	0	......	......	0	1	......	2
310–330	1	......	......	0	......	......	0	......	......	1
330–350	......	......	......	1	......	......	0	......	......	1
350–370	......	......	......	......	......	......	0	......	......	0
370–390	......	......	......	......	......	......	0	......	......	0
390–410	......	......	......	......	......	......	1	......	......	1
Total..	60	81	67	15	37	38	44	32	27	401

6. The following percentage distributions, columns 1 to 4 inclusive, are made up of relatives of weighted sub-groups of wholesale prices (10 groups, 41 sub-groups) on a 1926 base; that is, the 1926 price average for the sub-group is taken as 100%. Column 5 is similarly made up from individual price relatives (550 commodities) for the year 1927.

Plot each distribution (rectangular) and its cumulative. Note that the distributions are much less regular than similar distributions on a 1913 base.

Class mark, m	(1) 1923	(2) 1924	(3) 1925	(4) 1927	(5) 1927
55	....	0.8	....	....	0.1
60	0.8	0	....	....	0.2
65	0	0	....	....	1.4
70	0	0	....	....	3.7
75	9.8	8.7	....	8.5	6.5
80	6.7	6.7	....	0.7	7.3
85	7.0	8.1	....	0.3	5.9
90	3.4	0	....	5.5	12.0
95	2.2	8.7	20.3	35.1	13.9
100	19.7	32.8	33.2	36.2	25.9
105	13.6	9 5	25.6	11.8	9.8
110	10.3	11.7	5.9	.8	4.5
115	22 8	13 0	10.1	0	2.6
120	2 0	....	4.1	1.1	3.6
125	0.3	....	0	....	0.5
130	1.4	....	0	....	0.4
135	....	....	0	....	1.2
140	....	....	0	....	0.4
145	....	....	0	....	0
150	....	....	0.8	....	0
155	....	....	....	....	0.1

7. Classification of Iowa rural properties by value and assessment ratio (assessed value to sale value). Double frequency table.

Assessment ratio

Value (m)	$15\frac{1}{2}$	$25\frac{1}{2}$	$35\frac{1}{2}$	$45\frac{1}{2}$	$55\frac{1}{2}$	$65\frac{1}{2}$	$75\frac{1}{2}$	$85\frac{1}{2}$	$95\frac{1}{2}$	$105\frac{1}{2}$	$115\frac{1}{2}$	$125\frac{1}{2}$	$135\frac{1}{2}$
300	....	17	4	1	2	1							
260	1	15	56	3	3	0							
220	2	28	251	65	1	0	1						
180	1	7	147	277	34	8	0	1	0	0	0	0	1
140	1	6	65	301	295	71	6	2	0	0	1		
100	....	3	19	53	128	109	55	13	1	0	0		
60	....	1	1	10	14	19	11	8	4	1	1		

Total the frequencies according to each scale separately, and plot each distribution.

8. The following table represents 16 cities, denoted by letters, scored for certain social characteristics as indicated by Indexes A and B. Prepare a double frequency tabulation ($i = 10$; $m = 10, 20$, etc.).

City	Index A	Index B	City	Index A	Index B
A	31	11	I	21	22
B	10	10	J	39	20
C	48	31	K	30	19
D	41	18	L	42	30
E	19	9	M	28	22
F	31	29	N	29	21
G	21	11	O	38	29
H	32	19	P	20	19

ANSWERS

1. (a)

m	f	mf	Σf	Σf %
7	2	14	2	10
9	5	45	7	35
11	6	66	13	65
13	4	52	17	85
15	2	30	19	95
17	1	17	20	100

(b)

m	f	mf	Σf	Σf %
9	2	18	2	10
11	5	55	7	35
13	6	78	13	65
15	4	60	17	85
17	2	34	19	95
19	1	19	20	100

2. (a)

m	f	Σf
1	1	1
2	4	5
3	10	15

(b)

m	f
1	0
2	2
3	1

4	5	20
5	6	26
6	4	30
7	1	31
8	0	31
9	1	32

4	5
5	2
6	4
7	2
8	3
9	1
10	2

3. (a)

m	f	Σf
3	1	1
4	14	15
5	25	40
6	27	67
7	18	85
8	9	94
9	4	98
10	2	100

(b)

m	f	Σf
6	2	2
8	12	14
10	24	38
12	25	63
14	17	80
16	10	90
18	7	97
20	3	100

4. (a)

m	f	Σf
100	2	2
120	5	7
140	6	13
160	4	17
180	2	19
200	1	20

(b)

m	f	Σf
50	2	2
60	5	7
70	5	12
80	5	17
90	2	19
100	1	20

5. Cumulatives of total:

A. 1, 3, 14, 42, 101, 148, 212, 265, 306, 344, 368, 373, 376, 379, 380, 382.

B. 1, 1, 5, 15, 43, 88, 163, 230, 287, 325, 359, 368, 379, 381, 385, 385, 385, 385, 385, 386, 386, 387.

C. 3, 14, 38, 82, 146, 216, 291, 335, 367, 389, 394, 396, 398, 399, 400, 400, 400, 401.

6. Cumulatives of Column 5 (total):

0.1, 0.3, 1.7, 5.4, 11.9, 19.2, 25.1, 37.1, 51.0, 76.9, 86.7, 91.2, 93.8, 97.4, 97.9, 98.3, 99.5, 99.9, 99.9, 99.9, 100.0.

7. Value

m: 300, 260, 220, 180, 140, 100, 60.

f: 25, 78, 348, 476, 748, 381, 70.

Ratio:

m: 15.5, 25.5, 35.5, 45.5, 55.5, 65.5, 75.5, 85.5, 95.5, 105.5, 115.5.

f: 5, 77, 543, 710, 477, 208, 73, 24, 5, 1, 2.

8. *A*. $m = 10, 20, 30, 40, 50$
$f = 1, 4, 6, 4, 1$

B. $m = 10, 20, 30$
$f = 4, 8, 4$

CHAPTER III

AVERAGES

AFTER data have been tabulated in convenient form, the next step is usually the computation of measures which are typical or representative of the data or of changes in the data. In tabulations not involving a time element, the typical or representative number is some sort of an average. In a time series, a trend line representing the average direction of change may be computed. This chapter will deal with the principal forms of averages in common use.

Strictly speaking, an average is a number which may replace each of a set of items in a given situation without changing the result, or it is one which balances the deviations from it. For example, if three persons were receiving daily wages of $3, $4, and $5, respectively, the common average of $4 would make up the same total wage payment as the three separate wages taken together; or the deviations below and above $4 would balance (3–4 balances 5–4). Thus the average figure is true with respect to the aggregate and is useful in giving a general impression of a large array of data, but it loses sight of the variability of the figures and may sometimes be quite misleading if considered apart from the data themselves. In fact, sometimes the common average may have scarcely any significance at all, as in a case where a few very large incomes are averaged with a multitude of small incomes. The average may fall between the two groups and thus be typical of neither. Again, an average may give a false impression when groups having widely different variabilities are contrasted; for example, the average of 99, 100, and 101 is the same as the average of 1, 100, and 199, yet the two groups are not at all alike. Thus we may have groups alike in respect to their average but decidedly unlike with respect to other important characteristics. Averages are exceedingly useful for summary purposes, and for many purposes of calculation, but they are susceptible to much abuse, and should be used with this danger constantly in view.

The term "average" is generally assumed to imply the arithmetic mean (the sum of the items divided by the number of items), and in fact should be so interpreted when used without qualification. But

several other forms of the average are also important, and should be clearly distinguished. For example, a box 9 in. long, 6 in. wide, and 4 in. high would have an average dimension of 6 in., although the common average (arithmetic mean) is 6.3 in. The figure 6 may be checked by noting that $6 \times 6 \times 6 = 9 \times 6 \times 4$, hence the average, 6, may replace the given items of the data without affecting the volume enclosed. This type of average balances the ratio deviations above and below it (9/6 times 4/6 equals unity).

An average is obviously a measure of the central tendency of the items, and as such is often chosen by a criterion less strict than that suggested above. For example, in the United States during the year 1918, when the mean income received was \$1,543, the income most commonly received was about \$957. This type of average is called the mode, implying that it is the most usual or common measure in the distribution. In the rectangular graphs of distributions presented in the previous chapter it may be roughly located as falling between the limits of the class having the largest frequency. Again, during 1918 one-half of the incomes received in the United States were below \$1,140 and one-half were above. Hence this figure may also be regarded as a measure of central tendency, and may be called the median because of its central position in an "array" of the incomes received. That is, if all the incomes were listed in the order of their size, \$1,140 would appear half way down the list. These brief illustrations may be taken as an introduction to the many types of averages commonly met with, the computation and significance of which will now be considered more in detail.

The common average, or arithmetic mean (A, M, or AM).—As has already been explained, this is the number which, when substituted for the items (m) from which it is derived, will yield the same sum. The sum of the deviations of the items from it is zero ($\Sigma \overline{m - AM} = 0$). It is obtained by adding the items (m) and dividing by the number of items (n); that is, $AM = \Sigma m/n$. When the data are tabulated, the items or deviations must obviously be multiplied by the frequencies as a part of the process of summation.* The usual processes are

* An expression like $\Sigma fd/n + M_a$ is an abbreviation for $f_1d_1/n + f_2d_2/n + f_3d_3/n$, etc., plus M_a. The summation sign (Σ) governs as far as the next plus or minus sign. But a constant governed by the summation sign may obviously be removed and written as a coefficient of the completed summation, so that $\Sigma fd/n$ may be considered as $(1/n)(\Sigma fd)$, or $(\Sigma fd) \div n$. Care must be taken in handling summations in the transformation of formulas. For a discussion of this point see the Appendix. It should also be noted that the symbol f is not necessarily required after a summation sign, since it is understood as a part of the process of summating. In fact, it is better omitted as a rule, since in algebraic transformations it merely creates confusion.

briefly illustrated in Example 1, where use is made of a short-cut method in which a convenient estimate of the average is first assumed and then corrected by reference to the mean deviation of the data from it.

Example 1.—The arithmetic mean of untabulated and tabulated data.

(I) Untabulated data. $AM = \Sigma m/n$; or $AM = M_a + c$, where M_a is an assumed average, and $c = \Sigma d/n = \Sigma(m - M_a)/n$. Add $M_a + c$ algebraically; that is, with regard to the sign of c. If $M_a = AM$, $c = 0$.

(a) Average of 4, 5, and 9. $AM = \Sigma m/n = (4 + 5 + 9)/3 = 6$.

Or, (b) 1. Assume any convenient average (M_a) as 5.
2. Take deviations ($d = m - M_a$); $4 - 5 = -1$; $5 - 5 = 0$; and $9 - 5 = 4$.
3. Average these deviations to obtain a correction (c): $(-1 + 0 + 4)/3 = 1 = c$.
4. Take $AM = M_a + c = 5 + 1 = 6$. (Observe sign of c in adding.)

(II) Tabulated data. $AM = \Sigma fm/n$; or $AM = M_a + c$ where $c = \Sigma fd/n = \Sigma f(m - M_a)/n$. Add $M_a + c$ algebraically; that is, with regard to the sign of c. If $M_a = AM$, $c = 0$.

(a)

m	f	fm
3	2	6
5	4	20
7	3	21
9	1	9
	$n = 10$	$)56 = \Sigma fm$
		$AM = 5.6$

Or, (b)

m	f	d	fd
3	2	−2	−4
$M_a = 5$	4	0	0
7	3	2	6
9	1	4	4
			$10)6 = \Sigma fd$
			$c = 0.6$
			$M_a = 5.0$
			$AM = 5.6$

Certain observations may be made regarding the processes described in Example 1. In the first place, a further short cut may be made, in the writing of the deviations from an assumed mean in the case of tabulated data, by writing so-called step-deviations, in which the class interval (i) is taken as a unit. If this is done, the deviations (d/i) in Example 1 (II *b*) become -1; 0; 1; 2, which when multiplied by the frequencies give a total of $\Sigma fd/i = 3.0$; or $c/i = 0.3$; multiplying c/i

by i gives c; that is, $c = 0.3 \times 2 = 0.6$. This procedure is often a convenient short cut.

It may also be observed that in formulas involving frequencies it is not really necessary to express f, since multiplying by f is merely a part of the process of summation. That is, the expression Σd implies multiplying by the frequencies, if any are given. Also, Σf taken alone is by definition n.

Weighted averages.—A weighted average, where greater importance is attached to some magnitudes than others, follows the same form as that illustrated for tabulated data in Example 1. A weight simply indicates the number of times a given item is to be counted, and is therefore equivalent to a frequency. The sum of the weights, as of the frequencies, is therefore n. For example, if a term grade of 80 is to be averaged with an examination grade of 86, giving double weight to the former, the weights, 2 and 1, are taken as the respective frequencies, and the average is found as follows: $80 \times 2 + 86 \times 1 = 246$; which is divided by the sum of the frequencies, 3, to give the average 82. Several examples of weighted averages will appear below. Occasionally both frequencies and weights may be called for in the same averaging, in which case they are combined by multiplying the weight times the frequency for each class, and the product is treated as a revised frequency. Thus a measure (m) having a weight of 2 and a frequency of 5 would be treated as having a frequency of 10.

It is readily seen that in one sense all averages are really weighted averages, unit weights being understood where none are expressed. This fact explains certain apparent contradictions involved in averaging ratios, as appears in Example 2.

Example 2.—The average of prices and their reciprocals.

	Price per lb.	Amount per dollar
Bought 1st day..................	\$0.05 a lb. =	20 lb. for a dollar
Bought 2nd day..................	0.25 a lb. =	4 lb. for a dollar
	AM = \$0.15 a lb.	AM = 12 lb. for a dollar

But 15 cents a pound is not the same as 12 lb. for a dollar. In averaging 5 cents and 25 cents, weights of 1 lb. in each case are assumed, while in averaging 20 lb. and 4 lb. weights of one dollar in each case are assumed, and the two averages thus represent different transactions. To know the average price, we must know how much was actually sold. Suppose 40 lb. of each were sold, or \$2 worth the first day and \$10 worth the

second day; then consistent averages of reciprocals may be obtained by the appropriate use of consistent weights, as shown in Example 3.

Example 3.—The weighted averages of prices and their reciprocals, the amount per dollar (1 ÷ price). Prices are weighted by quantities purchased, and amounts per dollar by the number of dollars spent.*

	Price (p)	Weight	Product	Amount ($1/p$)	Weight	Product
Bought 1st day.....	\$0 05 ×	40 lb. =	\$ 2 00	20 lb. ×	\$ 2 00 =	40 lb.
Bought 2nd day.....	0.25 ×	40 lb. =	10.00	4 lb. ×	10 00 =	40 lb.
		80	80)12.00		\$12 00 =	12)80
		AM =	0.15			$6\frac{2}{3}$ lb.

In averaging prices per pound, the weight is the number of pounds purchased at a specified price; in averaging the number of pounds bought for a dollar (1 ÷ price) the weight is the number of dollars spent for each amount thus expressed. If the data correspond, the averages correspond (15 cents a pound is $6\frac{2}{3}$ lb for a dollar). So also in averaging any ratio number, the weight in cases where the common average is appropriate is the unit of reference, i.e., the denominator following "per." If no weight is expressed, a unit weight is implied. For example,

Number	Weight
Price per pound..........................	Pounds
Amount per dollar........................	Dollars
Persons per square mile....................	Square miles
Miles per hour...........................	Hours
Minutes per unit of work..................	Units of work
Interest rates, $\frac{¢}{100}$ per dollar...............	Dollars

In all averaging it is important to distinguish between ratios and the so-called fundamentals that make up the ratios. For example, price is the ratio of the fundamentals, dollars and quantities, and speed is the ratio of the distance traveled to the time elapsed. When the reciprocals of the ratios are employed, the price, speed, or other ratio is merely

* The averaging of ratios such as prices and amounts per dollar is most logically expressed as the ratio of the averages of the terms entering into the ratios, e.g., 6 ÷ 40 = 0.15, and 40 ÷ 6 = $6\frac{2}{3}$. Although this gives the same result as the ratio of the sums, 12 ÷ 80 and 80 ÷ 12, the distinction is important as leading up to the averaging of double ratios. The strict justification for calling 15 cents the average price is that it may be substituted for the given prices without changing average results; that is, the average price times the average quantity bought gives the average expenditure each day.

expressed in another form, e.g., miles per hour is changed to hours per mile. It is evident that the weights applied to ratios represent the fundamentals used as the denominator of the ratio, while in averaging the fundamentals themselves, no weights except unit weights are ordinarily required. However, cases may arise where the weights employed have no definite functional relation to the numbers averaged, but represent merely an estimate of the relative importance of the items.

Problems involving ratios and their reciprocals, like the price problem above, may conveniently be written in the accompanying form. The two fundamental series are written in the center, and the ratio series at the extremes, in such a way that the weights used in averaging the ratios are adjacent to them. Both forward and reverse, the columns read: Ratio, Weight, Product, Reciprocal of ratio (cf. Example 4).

Example 4.—Simultaneous weighted averages of reciprocals.

	(Ratio v/q) Price (p)	(Fundamental series) Quantity (q)	Value (v)	(Ratio, q/v) Amount ($a = 1/p$)
Bought 1st day....	\$0.05	40 lb.	\$ 2.00	20 lb. for \$1
Bought 2nd day....	0.25	40	10.00	4
	AM_w = \$0.15	Σ = 80	Σ = \$12.00	$AM_w = 6\frac{2}{3}$
		AM = 40	AM = 6.00	

$12 \div 80 = 0.15$, average price; $80 \div 12 = 6\frac{2}{3}$, average amount ($0.15 \times 6\frac{2}{3} = 1$)

The geometric mean (*G* **or** *GM*).—This is the number which when substituted for the items (m) from which it is derived will yield the same product. It is the nth root of the product of the n items ($G^n = \Pi m$).* It balances the ratio deviations of the items from it; that is, the product of the ratios of each item to it is unity ($m_1/G \times m_2/G$; etc., $= 1$). Thus the geometric mean of 4 and 9 is $(4 \times 9)^{1/2} = 6$; and of 5, 8, and 25 is $(5 \times 8 \times 25)^{1/3} = 10$. Usually the geometric mean is most easily found as the antilog of the average logs of the numbers (m), as here illustrated (cf. Example 5).

Example 5.—The geometric mean (G or GM). The logarithms of the items (m) are averaged, and the antilog of this average is taken as the geometric mean

* The symbol Π (capital letter) used before a variable term means the product of the variables. Thus Πm means m_1 times m_2 times m_3, etc. This symbol should not be confused with π (small letter) which indicates the ratio of the circumference to the diameter of a circle.

($GM^n = \Pi m$). In tabulated data the logarithms are weighted by the frequencies. For explanation and table of logarithms, see Appendix.

(a) Untabulated data.

m	$\log m$
2	0 3010
5	0.6990
10	1.0000
	3)2.0000
	$\log G = 0.6667$
	$G = 4.642$

(b) Tabulated data.

m	$\log m$	f	$f(\log m)$
3	0.4771	2	0.9542
5	0.6990	4	2.7960
7	0.8451	3	2.5353
9	0.9542	1	0.9542
		10	)7.2397
			$\log G = 0.72397$
			$G = 5.296$

Comparisons will show that the geometric mean is less than the arithmetic mean, and a mathematical proof for this relationship will be found in the Appendix (page 313). If applied to the prices and pounds for a dollar previously used, the geometric mean will give averages that are in harmony, as shown in Example 6.

Example 6.—The geometric means of reciprocals. The two means thus found will necessarily be reciprocals; for example, the geometric mean of a, b, and c is $\sqrt[3]{abc}$; and of $1/a$, $1/b$, and $1/c$, $1/\sqrt[3]{abc}$.

	Price	log	Amount	log
1st day.......	\$0.05	0.69897 − 2	20 lb.	1.30103
2nd day.......	0.25	1.39794 − 2	4 lb.	0.60206
		2)2.09691 − 4		2)1 90309
		$\log G = 1.04846 - 2$		$\log G = 0.95154$
		$G = 0.1118$		$G = 8.9442$

The two averages thus found are consistent; that is, a price of \$0.1118 is equivalent to 8.9442 lb. for a dollar ($0.1118 \times 8.9442 = 1$). Hence it was once thought that the geometric mean was the correct average to apply to ratios such as prices, amounts per dollar, and index numbers. But the geometric mean is in effect merely a weighted average. With two items each weight is the square root of the reciprocal of the item; for example, the weight for 0.05 is $(1/0.05)^{1/2} = 4.472$, and the weight for 0.25 is $(1/0.25)^{1/2} = 2$. When more than two prices are to be averaged, the case is more complex, but the principle is the same. The geometric mean therefore implies an arithmetic mean with assumed weights; these weights may constitute the best assumption possible if the real weights are not known; but the result is not the true average unless the weights in fact represent the true quantities and values

traded.* The geometric mean is not therefore of great importance in such cases. Its most important use is illustrated in processes involving multiplication, as in averaging successive rates (r) of increase of population over successive years or decades, where the geometric mean of the rates, each plus one, is taken as the average rate plus one $(1 + R)$, as illustrated in Example 7. The unit is added to the rates because $1 + r$ is the real factor by which the population is multiplied. The method may be applied to any number of successive increases. If a decrease is included, it is written as negative. A negative result will appear as a negative log of $(1 + R)$, and a fractional $(1 + R)$ which, taken less one, will indicate the average decline.

Example 7.—The average rate of population increase for two successive decades. The original population times $1.35 \times 1.08 \times 0.97 \times 1.22$ is the population at the end of the fourth decade; hence the average rate is the geometric mean of the rates, each plus one.

	Per cent increase or decrease (r)	$1 + r$	$\log (1 + r)$
1st decade.....	35%	1.35	0.13033
2nd decade....	8	1.08	0.03342
3rd decade....	−3	0.97	0.98677 − 1
4th decade....	22	1.22	0.08636

$$\begin{aligned} &1\ 23688 - 1 = \\ &4)0\ 23688 \\ \log (1 + R) &= 0.05922 \\ (1 + R) &= 1.1461 \\ R &= 0.1461. = 14.61\% \end{aligned}$$

Proof:

$$1.35 \times 1.08 \times 0.97 \times 1.22 = 1.7254.$$
$$1.1461 \times 1.1461 \times 1.1461 \times 1.1461 = 1.7254$$

that is, the constant rate, 14.61%, is here equivalent to the given rates, 35%, 8%, −3%, and 22%.

* When prices are expressed as ratios of the two fundamental terms implied in the transaction without reduction of the ratios to simpler forms, the average of the ratios is the sum of the numerators over the sum of the denominators. Thus if 144 lb. cost \$64 and 256 lb. cost \$100, the prices are respectively $\frac{64}{144}$ and $\frac{100}{256}$, the average of which is $(64 + 100)/(144 + 256) = 164/400 =$ \$0.41, as compared with an unweighted average of \$0.41$\frac{3}{4}$. But if one of these fractions is reduced, for example, if $\frac{64}{144}$ is written as $\frac{4}{9}$, the weighting is in effect changed and the above procedure will give the incorrect answer, \$0.39, owing to the lighter weighting of the ratio which has been reduced. But if the geometric mean is employed, giving the answer \$0.41$\frac{2}{3}$, this result is not changed if the ratios are reduced. The geometric mean of a series of prices (p) is the same as the geometric mean of the values (pq) divided by the geometric mean of the quantities (q) and is consistent with the geometric mean of the amount-per-dollar; that is, the geometric mean of the two fundamentals and the two ratios derived from them are consistent. This fact has a bearing on index numbers, to be discussed later.

Sometimes average rates of increase require weights, in which case a weighted geometric mean of the items, $1 + r$, is taken.

The harmonic mean (H or HM).—In averaging the reciprocals of prices-per-pound (that is, the amounts-per-dollar) it was seen that the unweighted mean implied that equal values of the commodities were traded. The result, transformed into the average price per pound, is called the harmonic mean of the given prices. Generally speaking, the harmonic mean is described as the reciprocal of the average of the reciprocals of the numbers (m). The average of the reciprocals is expressed as $\Sigma(1/m)/n$, and the reciprocal of this average, that is, the harmonic mean, may most conveniently be found as $HM = n/\Sigma(1/m)$.* If frequencies are present, they must be taken account of in the summation, as in other weighted averages. In this case the formula may be written, $HM = n/\Sigma(f/m)$; or if the frequencies are regarded as weights, $\Sigma w/\Sigma(w/m)$. The calculation of the harmonic mean of tabulated data assumes a form comparable to the geometric mean, in that the given measures are transferred to another scale (in the one case reciprocals and in the other case logarithms) for the purpose of averaging, and the average thus found is changed back to the original scale. For both untabulated and tabulated data, the procedure is as illustrated in Example 8.

Example 8.—The harmonic mean, $HM = n/\Sigma(1/m)$ or $n/\Sigma(f/m)$. The harmonic mean is the reciprocal of the average of the reciprocals of the given items. In extended tabulations, when calculating tables are used, it may be more convenient to write $1/m$ as illustrated. But sometimes it may be more convenient to omit $1/m$ and find f/m, which equals $f(1/m)$.

(a) Untabulated data.

m	$1/m$
0.05	20
0.25	4
	—

$HM = 2 \div 24 = 0.083$

(b) Tabulated data.

m	$1/m$	f	$f(1/m)$
3	0 333	2	0.666
5	0 200	4	0.800
7	0.143	3	0.429
9	0.111	1	0.111
		—	——

$HM = 10 \div 2.006 = 4.985$

* The unweighted harmonic mean of two numbers, a and b, is

$$HM = 2ab/(a + b)$$

and of three numbers, a, b, and c, is

$$HM = 3abc/(ab + ac + bc).$$

These formulas are easily obtained by algebraic transformation of the formula

$$HM = n/\Sigma(1/m).$$

The harmonic mean may be regarded as a form of the weighted average in which the weights consist of the reciprocals of the numbers ($w = 1/m$). If frequencies are present, the weights may be combined with the frequencies to form a new set of weights or frequencies ($w = f/m$). If the calculation is put down in this form it will parallel that given in Example 8 except that the order of the columns will be rearranged. Hence there is no particular advantage in the method except that it helps to emphasize the significance of the process. The nature of the weight implies that the harmonic mean must be less than the arithmetic mean, since in the former case the large numbers are given a relatively small weight and the small numbers a large weight, thus stressing the small numbers at the expense of the larger.

It should be noted that both the harmonic mean and the geometric mean are not ordinarily applicable to numbers which may vary to include zero. When the numbers to be averaged are very large relative to their variability, as 999, 1000, and 1001, the arithmetic, geometric, and harmonic means are almost identical, though the progression $AM > GM > HM$ will be found to hold. But if one of the series to be averaged is zero, as 0, 10, 20, both the geometric and harmonic means become zero, since the log of zero is minus infinity, and the reciprocal of zero is infinity. In each case, infinity renders the other items inoperative and determines the result.

Comparison of arithmetic and harmonic means.—In averaging ratios, it has been seen that the weights are normally in the units (U_d) implied in the denominator of the ratio. For example, prices-per-pound are weighted by pounds, and miles-per-hour by hours. But if the given weights are in the units (U_n) implied in the numerator of the ratio, then the harmonic mean is indicated. For example, prices-per-pound (p) if weighted by the number of dollars spent (v) take the harmonic form because this is in effect the same as weighting by the quantity bought ($q = v/p$).

The relation of the nature of the weights to the choice of arithmetic or harmonic mean may be summarized as follows:

1. When weights are given or assumed, use:
   AM if given weights are of U_d type,
   HM if given weights are of U_n type.
2. When no weights are given or assumed, the use of:
   AM assumes equal weights of U_d type, and
   HM assumes equal weights of U_n type.

The contrast of unweighted arithmetic and harmonic means may be illustrated as follows. Suppose that two men doing piece work are

timed respectively at 5 minutes and 8 minutes for a given task; then at this rate their average is 6.50 minutes (*AM*) if they perform the same number of tasks, but 6.15 minutes (*HM*) if they work the same length of time. It will be seen that the rate, minutes per task, has minutes as U_n and pieces as U_d. The arithmetic mean implies that the number of pieces (U_d) made by each worker is the same; the harmonic mean implies that the length of time put in by each worker is identical.

It should be noted that weights in general need not indicate the exact quantities, dollars or other units, as the case may be. It is sufficient if the ratios of the weights are known. For example, the average price of 100 lb. at 5 cents, and 50 lb. at 20 cents,

$$\overline{5 \times 100 + 20 \times 50}/150 = 10$$

is just as easily obtained if we know merely that the weights are 2 to 1 respectively $(\overline{5 \times 2 + 20 \times 1}/3 = 10)$. It is evident that the importance of the weights lies in their relative size rather than in their absolute size.

The quadratic mean (Q or QM).—Another average is the quadratic mean or root-mean-square, which is obtained by taking the square root of the average squares of the numbers; for example, the root-mean-square of 1, 7, 5, 3, and 4 is $\sqrt{(1 + 49 + 25 + 9 + 16) \div 5} = 4.47$. The common average of the same numbers is 4. But this is seldom applied to anything except sets of positive and negative deviations from some point of central tendency, in which case it is called the standard deviation. It will be illustrated later, in the chapter on dispersion.

Averages of position.—Two other forms of the average should be mentioned, which represent rather an estimate of central tendency than a strict mathematical concept, except in cases where the distribution is taken to represent a strict mathematical type. These measures are the median and the mode, sometimes called averages of position.

The median (*Md*).—In an array of items (the items listed in the order of their size), the median is simply the middle item, or the average of the two middle items; thus

(*a*) Numbers as given: 10, 7, 4, 15, 8. Numbers arrayed:
4, 7, 8, 10, 15; $Md = 8$.

(*b*) Numbers as given: 5, 15, 2, 7, 22, 9. Numbers arrayed:
2, 5, 7, 9, 15, 22; $Md = \frac{1}{2}(7 + 9) = 8$.

Thus the median may be described as the actual or interpolated item in the middle position of the items arranged in order of size. It is therefore not affected by the size of the extreme items, and is particularly useful where such items are unrepresentative or unreliable. In an array

the middle position is easily located by counting the items consecutively until the item numbered $(n + 1)/2$ is reached. Or, if more convenient, the spaces between the items may be counted until the space numbered $n/2$ is reached. Thus, in the first illustration (*a*) above, the median is located as the third item $(\overline{n + 1}/2 = 3)$; in the second illustration (*b*) the median is located as the numbers adjacent to the third space $(n/2 = 3)$. The relation of the median to the mean is illustrated in Chart 7.

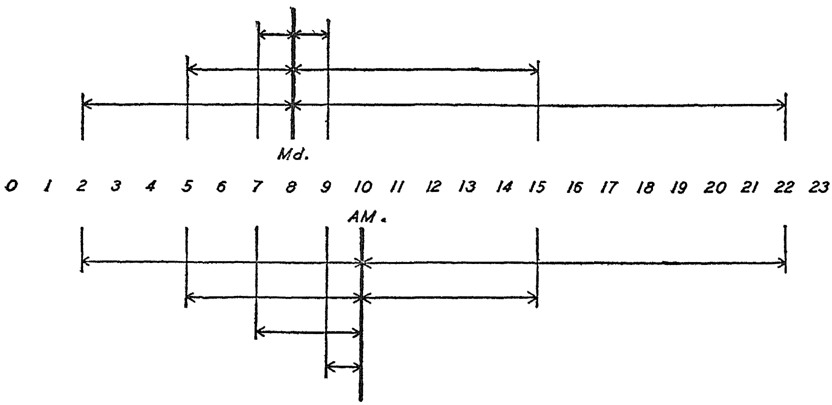


CHART 7

The relation of the median (*Md*) to the arithmetic mean (*AM*): $m = 2, 5, 7, 9, 15, 22$; $Md = 8$, $AM = 10$. The sum of the deviations from *Md* is 10 on the negative side and 22 on the positive side, a total of 32; the deviations from *AM* are 17 on the negative side and 17 on the positive side, a total of 34. The deviations about the median are always a minimum; that is, no other origin than the median, or the space in which the median occurs, will give a smaller total of deviations. The sums of the deviations from *AM* always balance; that is, their algebraic sum is zero.

The median of a frequency tabulation.—In a frequency tabulation, the median may be interpolated on the assumption that the magnitude scale is "continuous"; that is, that it may be interpolated at any point. This assumes that the frequencies are taken as representative of a very large group similarly distributed. The process is shown in Example 9.

Example 9.—Interpolating the median: $Md = L_1 + i(n/2 - \Sigma_1)/f$. The specific notation in the formula refers to the median class.

Class mark m	Classes L_1 L_2		Frequency f	Cumulatives Σ_1	Σ_2	
3	\$2–\$4		20	0	20	
5	4– 6	(median class)	40	20	60	(includes $n/2$)
7	6– 8		30	60	90	
9	8–10		10	90	100	
	$i = 2$		$n = 100$			
			$n/2 = 50$th space			

$$Md = L_1 + i(n/2 - \Sigma_1)/f = 4 + 2(50 - 20)/40 = 4 + 1.50 = \$5.50.$$

Since the class limits are theoretically at spaces between items, it is necessary to locate the middle space, which is the $n/2 = 50$th space. This obviously lies in the second class, between the cumulatives 20 and 60; and the median is therefore a point to be interpolated in the magnitude scale between the limits corresponding to 20 and 60, or between \$4 and \$6, as is done by the formula $Md = L_1 + i(n/2 - \Sigma_1)/f$. The

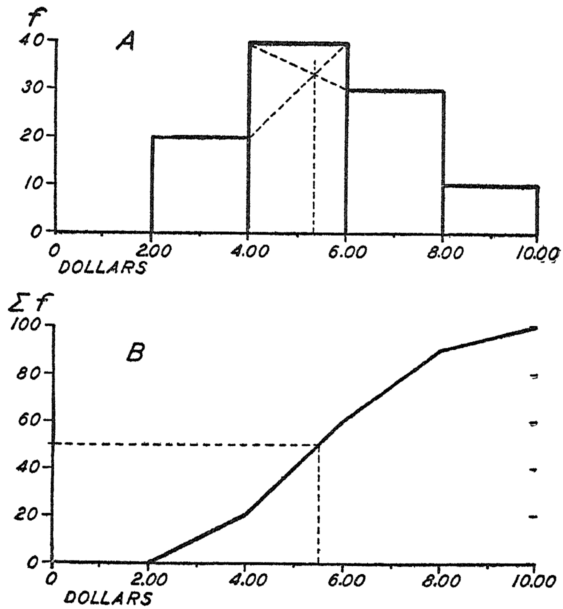


CHART 8

Graphic interpolation of the median and mode. Data of Examples 9 and 10 plotted as a rectangular frequency distribution (above) and as a "less than" cumulative curve (below). In the upper figure the mode is interpolated by drawing diagonal lines from the upper corners of the modal rectangle to the nearest upper corners of the adjacent frequency as indicated. The point of intersection of these diagonals marks the ordinate of the mode, and a perpendicular dropped from the intersection marks the mode on the magnitude scale. The median is interpolated in the cumulative curve (below), by drawing a horizontal line from $n/2$ (vertical scale) until it intersects the curve, and at the point of intersection dropping a perpendicular to the magnitude scale. The foot of this perpendicular indicates the median. The graphic interpolation of the median and mode thus described should check with the results calculated by the methods of Examples 9 and 10.

form may be abbreviated by writing simply the columns L_2 and Σ_2. This interpolation assumes that in each class the items are scattered at equal spaces apart, with half a space between each class limit and the adjacent item. In reality, however, the items in each class tend to cluster thicker toward the greatest frequency (more precisely, the mode), but the assumption is accurate enough for most purposes. A

more accurate method of interpolating the median will be discussed in the next chapter.

The median may usually be obtained with sufficient accuracy for practical purposes by interpolation in a chart of the cumulative curve as illustrated in Chart 8. The cumulative curve is drawn as previously described in Chapter II, but the rectangular frequencies are dropped as being unnecessary for the purpose at hand. It should be observed that the upper cumulatives, Σ_2, are plotted against the upper limits (L_2) of their respective classes, beginning with the zero of the curve at the lower limit of the initial class (the class having the smallest class mark). It is a common error to plot these cumulatives at the middle of the class, but it is obvious from previous charts that this does not represent the cumulatives correctly. As has previously been noted, the cumulative curve thus drawn is designated the " less than " summation, since any point on it designates the number of workers (vertical scale) receiving less than a given wage (horizontal scale). The curve may also be drawn as a " more than " cumulative by beginning the summation at the opposite end of the tabulation.

The Mode (*Mo*).—The mode is the most common actual or interpolated magnitude. It is usually obtained from a frequency table, either broadly by stating the class having the greatest frequency, or by interpolation if, as in the case of the median interpolation, the scale may be considered continuous and the frequencies representative. In the frequency distribution of Example 10, the mode is broadly the second class; that is, more of the wage earners in question fall in this class than in any other class. Obviously, the mode is limited to distributions that are somewhat regular. Hence the mode cannot be so widely used as the median, but it is even less influenced by erratic and extreme items.

The mode as interpolated is at best an estimate.* It is most exactly

* The method of interpolating the mode here explained is valid only as an approximation for smooth normal distributions. Hence its use is limited, but at least it serves the purpose of introducing the concept of interpolation. For positively skewed logarithmic distributions it usually gives too high a value, which may be roughly corrected by the following process. Write log m of the modal class and of the preceding and following classes, and take the differences of these logs in sequence. In the example given below, these logs are: $\log 3 = 0.4771$; $\log 5 = 0.6990$; $\log 7 = 0.8451$, and their first differences are 0.2219 and 0.1461. Next divide d_1 by the first of these logarithmic differences and d_2 by the second. That is, take $20/0.2219 = 90.13$ and $10/0.1461 = 68.45$, and use these quotients as revised d_1 and d_2, respectively. The formula for the mode will now read: $Mo = 4 + 2 \times 90.13/(90.13 + 68.45) = 5.14$. Since the given distribution is not strictly logarithmic this correction is a little too large. If, however, modes are to be interpolated more precisely, the method explained later (p. 304) may be used

approximated by curve fitting, to be discussed later. However, in fairly regular distributions it may be placed by dividing the modal class into parts which are proportional to the differences (d_1 and d_2) between the modal frequency and the adjacent frequencies, respectively, as shown in Example 10.

Example 10.—Interpolating the mode: $Mo = L_1 + id_1/(d_1 + d_2)$, where L_1 refers to the modal class, and d_1 and d_2 refer to the differences between the modal frequency (f_m) and the preceding (f_{-1}) and the following (f_{+1}) frequencies, respectively; that is, $d_1 = f_m - f_{-1}$, and $d_2 = f_m - f_{+1}$.

Class L_1 L_2			Frequency f	
2– 4		f_{-1}	20	$d_1 = 40 - 20 = 20$
4– 6	(modal class)	f_m	40	
6– 8		f_{+1}	30	$d_2 = 40 - 30 = 10$
8–10			10	
$i = 2$		$n =$	100	

$$\begin{aligned} Mo &= L_1 + id_1/(d_1 + d_2) \\ &= 4 + 2 \times \tfrac{20}{30} = 4 + 1\tfrac{1}{3} = 5\tfrac{1}{3}. \end{aligned}$$

The interpolation is sometimes based upon the adjacent frequencies rather than upon the differences, as above. But the method here given is preferable in that it finds the apex of a parabola passing through the three central frequencies, and it adjusts more consistently the variability of the mode to the variability of the frequencies. If two classes in the modal position have exactly the same frequencies, the common class limit may be regarded as the mode.*

The mode, median, and average are commonly given in describing a distribution. If these three averages are practically alike, the distribution probably approaches the normal. If the average is considerably greater than the mode, the distribution is positively skewed; that is, the curve slopes more gradually to the right than to the left. In such distributions, the median (Md) is between the mode (Mo) and the average (AM); roughly, $Mo = 3Md - 2AM$; or more strictly,

($Mo = 5.28$) or graduating formulas may be employed (cf. *Proceedings of the Casualty Actuarial and Statistical Society of America*, Vol. VI, Part I, pp. 52–72, or Reitz, H. L., "The Handbook of Mathematical Statistics," pp. 112–3.)

* The mode may be readily interpolated by the graphic method, as illustrated in Chart 8. Lines are drawn from the upper corners of the modal frequency rectangle diagonally to the nearest corners of the adjacent frequency rectangles. The point of intersection of these lines is on the modal ordinate; that is, the mode may be read on the X-scale perpendicularly below the point of intersection. The reading thus made theoretically agrees with the result obtained by the method of Example 10, as may easily be proved by elementary geometry.

$\log Mo = 3 \log Md - 2 \log AM$, if the distribution is a logarithmic normal.

Variations in percentile interpolations.—In linear interpolations of the median or other percentiles in tabulated data, it is sometimes convenient to cumulate the frequencies in reverse order, that is, beginning with the classes having the largest magnitudes and proceeding toward those having the lowest magnitudes. When this is done the linear interpolation formula given in the text will obviously require modification. Whether the cumulatives are written forward or reverse, it is also possible to write the formulas to indicate the interpolation from the lower or the upper limits of the class indicated. Thus four linear formulas may be written, two of them referring to forward cumulatives ("less than"), and two of them referring to reverse cumulatives ("more than"). In writing the reverse cumulatives it is well to reverse the position of the columns Σ_2 and Σ_1 so that they correspond to L_1 and L_2, respectively. A brief tabulation with the four formulas is given in Example 11.

Example 11.—Interpolation of the median from "less than" cumulatives (1) written forward, and "more than" cumulatives (2) written in reverse order. Also, in each case, the interpolation is made (a) from the lower limit (L_1) and (b) from the upper limit (L_2) of the median class (MdC).

			(1) Forward	(2) Reverse	
	L_1 L_2	f	Σ_1 Σ_2	Σ_2 Σ_1	
	2– 4	2	0– 2	10–8	
MdC	4– 6	4	2– 6	8–4	(includes $n/2$)
	6– 8	3	6– 9	4–1	
	8–10	1	9–10	1–0	
		$n = 10$			

(1) Cumulatives forward.

(a) Interpolating from L_1:

$$Md = L_1 + i(n/2 - \Sigma_1)/f = 4 + 2(5 - 2)/4 = 5.5.$$

(b) Interpolating from L_2:

$$Md = L_2 - i(\Sigma_2 - n/2)/f = 6 - 2(6 - 5)/4 = 5.5.$$

(2) Cumulatives reverse.

(a) Interpolating from L_1:

$$Md = L_1 + i(\Sigma_2 - n/2)/f = 4 + 2(8 - 5)/4 = 5.5.$$

(b) Interpolating from L_2:

$$Md = L_2 - i(n/2 - \Sigma_1)/f = 6 - 2(5 - 4)/4 = 5.5.$$

SUPPLEMENTARY METHODS

Comparisons of averages.—The question sometimes arises whether it is better to use the average of ratios, or the ratio of the averages of the series of items which make up the ratios, as illustrated in the accompanying figures for bank debits in eight Iowa cities, September, 1928 and 1929 (see Example 12).

Example 12.—Averaging the ratio of change.

Bank debits in thousands of dollars

City	(m) Ratio, 1929/1928	September	
		1928 (weights)	1929 (products)
Cedar Rapids	114.5%	44,979	51,497
Davenport	106.1	41,616	44,134
Des Moines	108 5	79,343	86,087
Dubuque	98.7	15,925	15,725
Mason City	115 2	11,090	12,773
Muscatine	94.6	5,821	5,506
Sioux City	98.5	60,472	59,550
Waterloo	100.3	22,692	22,767
Totals	8)836 4	8)281,938	8)298,039
Averages	104.6	35,242	37,255
Weighted average	105.7		

In this problem the unweighted average of the ratios obviously fails to make allowance for the varying importance of the cities. But the weighted average of the ratios (105.7), obtained by using the debits in the base year, 1928, as the weights, is identical with the ratio of the averages for each year (i.e., 37,255 ÷ 35,242 = 105.7). This weighted average is also the same as the ratio of the totals for each year. Hence, this is a consistent average of the rate of change. But if the data are considered mere samples, more arbitrary weights in harmony with the assumed representative character of the cities might be employed.

If the data were incomplete, the ratios but not the figures from which they were computed being given, then a strict average would require weights representing the relative normal amount of debits in each city. If the average bank clearings in each city for recent years were known, these figures might be used as weights (w). Since they would be comparable, relatively, to the combined two items from which the ratios (m) were derived, they should preferably be adjusted

as either $w/(m + 1)$ or $w/m^{1/2}$ (cf. Example 13). The resulting averages may be called the unit harmonic and root harmonic, respectively. These two methods of averaging give nearly the same results; the former assumes that the common average of the debits in 1928–1929 is proportional to the clearings, whereas the latter assumes that the geometric mean is proportional. The weighted geometric mean would also give nearly the same result, and is most commonly used, but its significance cannot be stated except in abstract formulas.*

Example 13.—Illustration of the unit-harmonic mean: $UH = \Sigma(mw \div \overline{1 + m}) \div \Sigma(w \div \overline{1 + m})$, and the root-harmonic mean: $RH = \Sigma(mw/m^{1/2}) \div \Sigma(w/m^{1/2})$, where $m = b/a$ is a series of positive variable numbers, and the original weights (w), which are proportional to $a + b$ and $a^{1/2}b^{1/2}$, respectively, are modified in use as indicated by the formulas.

Assume bank debits, a and b (millions of dollars), in three cities at successive dates, and the ratios forward and backward.

City	$m = b/a$	a	b	a/b
A.......	250%	180	450	40%
B.......	200	125	250	50
C.......	125	120	150	80
Average	200%	425	850	50%

I. Assume that the only data are b/a and the weights $w = a + b = 630$, 375, and 270, respectively. Find weighted average of b/a using as actual weights $w_a = w \div (1 + m)$.

$m = b/a$	$w_a = w \div (1 + m)$	Product
2.50	(630/3.50 = 180)	450
2.00	(375/3.00 = 125)	250
1.25	(270/2.25 = 120)	150
Average..2.00	425	)850
		2.00 = Average b/a

Hence if weights are given approximating the mean of the fundamentals, the unit harmonic mean of the ratios is the correct mean.

II. Assume again b/a, and the weights $w = a^{1/2}b^{1/2} = 284.60$, 176.78, and 134.16, respectively. Find weighted averages of b/a, using as actual weights $w_a = w/m^{1/2}$.

* For two numbers weighted, as a weighted n_1 times and ab weighted n_2 times, where $n_1 + n_2 = n$, the weights are as follows:

No. (m)	Weight (w)
a	$a^{-0.5}\ (b^{(n_1-1)/2n} + b^{(n_1-3)/2n} + \ldots b^{(1-n_1)/2n})$
ab	$a^{-0.5}\ (b^{-(n_1+1)/2n} + b^{-(n_1+3)/2n} + \ldots b^{(1-n_2-n)/2n})$

If $n_1 = n_2$, these weights are equivalent to weights of $1/m^{1/2}$.

By successive averaging, the geometric mean of any number of appropriate items ($m > 0$) may thus be found.

$m = b/a$	$w_a = w/m^{1/2}$	Product
2.50	(284.60/1.5811 = 180)	450
2 00	(176.78/1 4142 = 125)	250
1.25	(134.16/1.1180 = 120)	150
Average. . 2.00	425	)850
		2.00 = Average b/a

Hence if weights are given approximating the geometric mean of the fundamentals, the root harmonic mean of the ratios is the correct mean.

Averaging double ratios.—In deriving index numbers as discussed later it is required to take averages of ratios consisting of prices at one date compared respectively with prices at another date considered as the base. Such comparisons are called price relatives, and they are examples of double ratios (i.e., one ratio divided by another). The problem in its theoretical aspects is too complex to consider at this point, but it may be well to introduce a parallel but simpler problem in the averaging of double ratios.

Let us assume that two buyers, A and B, dealing in a given commodity make certain purchases during two consecutive weeks. During the first week A pays \$2.00 per unit bought, and during the second week \$3.00. On the other hand B, buying the same goods in another vicinity, pays \$5.00 per unit the first week and \$2.00 the second week. The percentages or price relatives, representing prices (p_1) in the second week as compared with prices (p_0) in the first week, are:

Price relative for $A = p_1/p_0 = 3/2 = 150$ (%)
Price relative for $B = p_1/p_0 = 2/5 = 40$ (%)

and the average of the relatives 150 and 40 is required.

Obviously a consistent average involves weightings with respect to the fundamentals of the ratios, that is, the goods and money changing hands. On the basis of limited data, the procedure is illustrated in Example 14.

Example 14.—Averaging double ratios. Buyers A and B purchase q at p prices ($pq = v$) during first or base week (wk_0) and second or given week (wk_1), respectively. The percentages below are the price, quantity, and value relatives as indicated. Averages of fundamentals, q and v, are unweighted; and of ratios, weighted. Averages in parentheses give a secondary solution.

	First week			Second week			Percentages		
	p_0	q_0	v_0	p_1	q_1	v_1	p_1/p_0	q_1/q_0	v_1/v_0
A.	2	5	10	3	8	24	150	160	240
B.	5	3	15	2	8	16	40	267	107
Averages. .	(3.125)	(4)	($12\frac{1}{2}$)	(2.50)	(8)	(20)	80	200	160

Average $q\%$ = average 160% and 267% (weights 5 and 3) = 200%.
Average $v\%$ = average 240% and 107% (weights 10 and 15) = 160%.
Average $p\%$ = average $v\%$/average $q\%$ = 160/200 = 80%.

Percentages based on the averages of the first and second weeks give the same relatives obtained above.

It should be observed, however, that the average of the double ratios is not always taken to imply full weightings as indicated by the situation itself. The reason for such a limitation is that the average of the double ratios when fully weighted in terms of the actual situation may, in extreme cases, give an average which lies outside the limits of variability of the double ratios themselves. Hence such weighted averages involve a broader meaning of the term *average* than is usually considered justifiable. Nevertheless, they may be regarded as consistent averages, since they sum up the change involved in the whole situation as against the change involved in each of the parts, and since also their computation involves the same logic as that of single ratios. Thus, it is quite possible that the records of a group of runners on successive days should all show a decline and yet the average rate of running would show an increase, since the running at the later date might be largely transferred from the slower to the speedier runners. This is an important point with respect to the theory of index numbers, as will appear later. It is conceivable that in extreme cases all individual prices might fall, and yet the price level might rise.

From Example 14 it may be seen that the average of price relatives (double ratios: $\overline{v_1/q_1} \div \overline{v_0/q_0}$) is readily obtained by carrying the usual averaging of ratios one step farther. The procedure requires that the q percentages for A and B must be weighted, just as the average of simple prices must be weighted. According to the usual procedure the required mean is the arithmetic mean weighted with q_0, or the harmonic weighted with q_1. These two means necessarily give identical results. The mean of the v percentages is obtained in a like manner. The average of the p percentages is then given as the ratio of the v and q averages, just as in any averaging of prices. The results check with the percentages obtained directly from the averages for the base and given weeks; that is, $2.50/3.125 = 80\%$; $8/4 = 200\%$ and $20/12.5 = 160\%$. Just as the average price is the ratio of the average fundamentals, so the average of price relatives is the ratio of the average prices. In general, the average of ratios, when consistently weighted, becomes identical with the ratio of the averages.

The process of averaging ratios and double ratios in the form of prices may be simply expressed thus:

$$\text{Ratio:} \qquad \text{Average } p = \Sigma v/\Sigma q.$$

$$\text{Double ratio: Average } p_1/p_0 = (\Sigma v_1/\Sigma v_0) \div (\Sigma q_1/\Sigma q_0).$$

The same method may be applied to the averaging of double ratios in other fields. For example, the speed of two cars expressed in miles-

per-hour during a run of two successive days might be tabulated in the same form as that used in Example 14, the fundamentals being hours and miles.

The problem of averaging price ratios when quantities of incommensurable units (pounds, yards, kilowatt-hours, etc.) are involved must be referred to the chapter on index numbers.

The average height of a curve.—If data are available such that over a given period of time the general trend of an index series (for example, production, total immigration, death rates, or other data) may be expressed as a mathematical curve of the parabolic type, the average of the series during the whole period may be calculated by a process involving the integration of the curve.

As an illustration, the following data may be taken. The time interval is 1921 to 1929 inclusive, represented by an x-scale on which the middle year, 1925, is taken as $x = 0$, and therefore 1921 is $x = -4$ and 1929 is $x = +4$. The equation of the curve is assumed to be of the general type, $T = a + bx + cx^2$, which in this particular case is taken as $T = 80 + 12x + x^2$. The curve may be plotted by solving the equation for each value of x from -4 to $+4$. Thus in 1921, where $x = -4$, the curve has the ordinate, or height

$$T = 80 + 12(-4) + (-4)^2 = 48$$

In 1922, where $x = -3$, the ordinate of the curve is:

$$T = 80 + 12(-3) + (-3)^2 = 53$$

If this process is carried out for each year up to 1929, the ordinates of the curve from 1921 ($x = -4$) to 1929 ($x = +4$) are found to be $T = 48, 53, 60, 69, 80, 93, 108, 125, 144$ (cf. Chart 9).

The average height of the curve may obviously be found as the area under the curve divided by the base. The curve and its base must be taken from January 1, 1921 ($x = -4.5$), to December 31, 1929 ($x = +4.5$), or 9 years. The area (A) is expressed in mathematical notation by the following so-called integral

$$A = \int_{-4.5}^{4.5} (80 + 12x + x^2)dx$$

which simply means that the area under the curve, $T = 80 + 12x + x^2$, is to be summed up between the limits -4.5 and $+4.5$ along the x-scale, as indicated by the final expression dx (d is not a factor but a symbol in calculus indicating the scale on which successive small increments are taken).

The rule for integrating any term in an equation of the parabolic curve is expressed thus: Let the term be represented by kx^n, where k may be any positive or negative constant, x is a point on the time scale as rearranged for purposes of computation, and n is a power of x usually ranging from 0 ($x^0 = 1$) to not more than 5 or 6. Then the area or integration of this term is expressed as $\frac{k}{n+1}x^{n+1}$; for example, the term 80 (times $x^0 = 1$, understood) integrates as $80x$; $12x$ integrates as

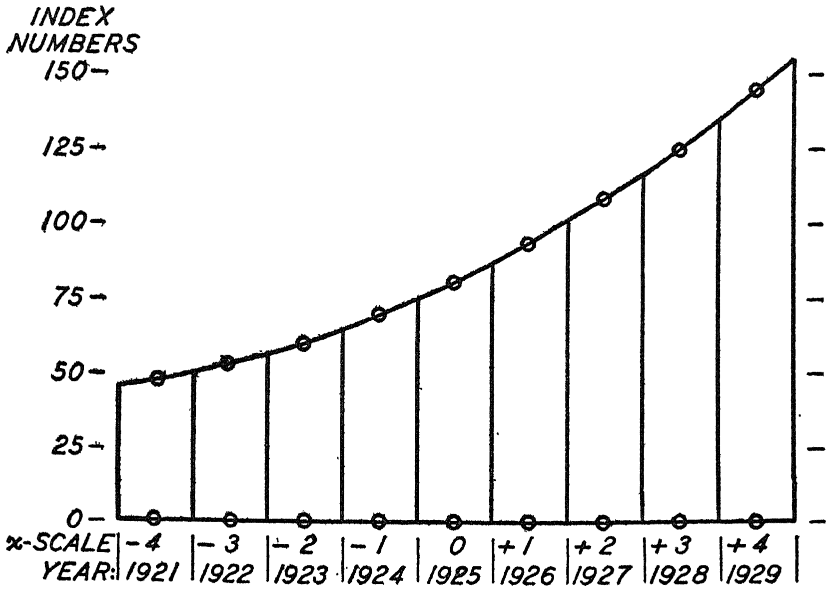


CHART 9

Averaging the height of a parabolic curve. The curve $T = 80 + 12x + x^2$; 1921–1929; origin, 1925; plotted as a smooth curve. The average, taking each annual ordinate as a rectangle, is

$$(48 + 53 + 60 + 69 + 80 + 93 + 108 + 125 + 144) \div 9 = 86.7.$$

The average of the smooth curve obtained by integration is 86.8.

$6x^2$, and x^2 as $\frac{1}{3}x^3$. Hence the general integral (area $= A$) of the equation $T = 80 + 12x + x^2$ is $80x + 6x^2 + \frac{1}{3}x^3$. This integral expresses the area under the curve from the origin ($x = 0$) to any specified x. It may theoretically include an undetermined constant which has no significance here.

If, next, the limiting values of x are substituted successively in this integral equation, the areas under the curve from the origin to the beginning of the curve ($x = -4.5$) and from the origin to the end of the curve ($x = +4.5$) are obtained. The unit of area in this case is a

composite, but it may be tolerated since it is not a final product. The sum of the two areas thus obtained is obviously the total area under the curve. Or, in general, the total area may be expressed as the final integral less the initial integral—a rule which applies both in the given case and in any case when the origin ($x = 0$) is outside the limits of the given time interval.

The required areas are as follows:

$$\text{When } x = -4.5,\ A = 80(-4.5) + 6(-4.5)^2 + \tfrac{1}{3}(-4.5)^3 = -268.9$$

The area in this case appears to be negative because the base is expressed in negative numbers. The area at the end of the interval is

$$\text{When } x = +4.5,\ A = 80(4.5) + 6(4.5)^2 + \tfrac{1}{3}(4.5)^3 = 511.9$$

The total area under the curve is therefore $511.9 - (-268.9) = 780.8$. The total area is now regarded as positive, being taken, in effect, on a time scale with a prior origin. This area, divided by its base, namely, the interval of time from $x = -4.5$ to $x = +4.5$, or 9 years, is 86.8. This is the average height. It could, of course, have been found approximately by averaging the ordinates of the successive x's from 1921 to 1929. This average is

$$(48 + 53 + 60 + 69 + 80 + 93 + 108 + 125 + 144) \div 9 = 780 \div 9 = 86.7$$

However, the rule by integration may be applied directly to the curve without calculating the successive ordinates, and besides it has many other uses. It is worth mastering as a simple introduction to one of the methods of the calculus.

It is sometimes desirable in problems involving a curve expressed as an equation to find the slope of the curve at any given ordinate on the time scale (x). By the slope is meant the increase per time unit of a line tangent to the curve at the given point of time. This slope may be found by a process called differentiation, which is the reverse of the process of integration just described. The result obtained by differentiating a trend equation (T) is called the derivative, and is denoted mathematically by the expression dT/dx, which is interpreted to mean the rate of rise or fall in the trend during a very small interval of time, dx. This ratio obviously expresses the slope or rise per unit at any given point, that is, the rate of growth then prevailing. The terms in an equation of a parabolic curve may be differentiated separately. Thus, denoting any term as kx^n (notation as before) the derivative is:

$$\frac{dT}{dx} = knx^{n-1}$$

Comparison with the formula previously given for integrating will show that differentiation is merely the opposite of integration. The rule for differentiating applied to the equation $T = 80 + 12x + x^2$ gives $dT/dx = 12 + 2x$. The term 80 (where $x^0 = 1$, understood) obviously drops out of the derivative equation. The derivative thus found will give the slope of the curve at any ordinate of x by substituting the required value of x; for example, at the beginning of the interval where $x = -4.5$ the slope is $dT/dx = 12 + 2x = 12 + 2(-4.5) = 3$; that is, at the beginning of the period the rate of growth is 3 units per year. At the middle of the period where $x = 0$ the rate is $12 + 2 \times (0) = 12$, and at the end of the period the rate is $12 + 2 \times (4.5) = 21$ units per year.

In interpreting curves such as the one just discussed, it is often desirable to note the time when the curve reaches its lowest point, if it is concave, or its highest point, if it is convex. The time may be found by noting that the slope then is obviously zero, and the equation for the slope may therefore be equated to zero thus:

$$\frac{dT}{dx} = 12 + 2x = 0$$

Solving the equation $12 + 2x = 0$ for x, gives $x = -6$. This means that at the point on the time scale where $x = -6$, or the year 1919, the curve reaches its lowest point.

EXERCISES

1. (I) Define the arithmetic, geometric, and harmonic means.
(II) Find the arithmetic mean by inspection:

(a)	5;	10;	15;	20;	25	
(b)	0;	2;	4;	6;	8	
(c)	−4;	0;	4;	8;	12	

(*Note.*—The mean of an arithmetic series, where n is odd, is the middle term.)

(d)	4;	8;	12;	16		
(e)	5;	10;	15;	20;	25;	30
(f)	−6;	−2;	2;	6;	10;	14
(g)	−6;	−4;	−2;	0;	2;	4

(*Note.*—The mean of an arithmetic series, where n is even, is the mean of the two middle items.)

(h)	1;	8;	9;	12;	15
(i)	6;	8;	9;	12;	15
(j)	6;	8;	9;	12;	150
(k)	$\frac{1}{10}$;	$\frac{1}{5}$;	$\frac{1}{4}$;	$\frac{1}{2}$;	1

(*Note.*—Change fractions to decimals.)

(l)	$\frac{1}{7}$;	$\frac{1}{3}$;	$\frac{2}{3}$;	$1\frac{1}{2}$	(to hundredths)

2. (a) Find the average (AM) net profits per quarter for the representative companies included below:

Net profits, 355 companies:	Millions of dollars
1927, 1st quarter	468
" 2nd quarter	520
" 3rd quarter	576
" 4th quarter	457

(b) Find the average (AM) building contracts awarded in the United States during the years 1920 to 1927 inclusive from the following record:

Year	Thousands of square feet	Year	Thousands of square feet
1920	459,300	1924	706,428
1921	442,308	1925	899,460
1922	654,624	1926	842,940
1923	676,224	1927	812,388

(c) Find the average (AM) of the following wage distribution:

Wages m	Workers f	Wages m	Workers f
\$2	50	\$8	40
4	80	10	20
6	60		

(d) Find the average (AM) of the following wage distribution:

Wages m	Workers f	Wages m	Workers f
\$2	20	\$5	80
3	40	6	50
4	60		

3. (*a*) Assuming an average of \$6.00, find the corrected average (AM) of the following tabulation:

Wages	f	Wages	f
\$2	1	\$6	3
4	5	8	1

(b) Recompute the foregoing, assuming an average of 5; of 7.

4. (a) Compute the average (AM) of the following five-class normal distribution, using successively the class marks as the assumed averages:

Wages	f	Wages	f
\$2	1	\$8	4
4	4	10	1
6	6		

(*Note.*—The average of a normal distribution lies at the middle of the series of class marks.)

(b) What is the average of a normal wage distribution of six classes, whose class marks are (in dollars): 2; 3; 4; 5; 6; 7?

5. (a) Find the average (AM) price per pound, assuming that equal quantities were sold at each price: Price per pound $0.50; $0.25; $0.80; $0.40; $1.25.
   (b) Recompute the average, assuming that equal values were sold at each price.
   (c) Find the average amount-per-dollar, assuming equal values sold: Pounds per dollar: 2; 4; $1\frac{1}{4}$; $2\frac{1}{2}$; 0.8.
   (d) Recompute (c) assuming that equal quantities were sold.
   (e) Compare the four preceding answers and explain.

6. (a) Find the average (AM) of 25 and 40 miles per hour, assuming that each speed was maintained for an equal length of time.
   (b) Recompute the average, assuming that each speed was maintained over an equal distance.
   (c) If three piece workers ordinarily perform a given task in 10, 12, and 15 minutes, respectively, what is the normal average time per task for these workers in a day?

7. Find the geometric mean, and compare with the arithmetic and harmonic means:

(a)	4;	10			
(b)	50;	80;	100;	200	
(c)	4;	16;	64;	256	
(d)	1;	10;	100;	1000;	10,000

(*Note.*—The GM of a geometric series, where n is odd, is the middle item of the series; and where n is even, it is the geometric mean of the two middle items.)

(e)	4;	9;		
(f)	8;	27;	125	
(g)	16;	81;	256;	625

(*Note.*—The GM may sometimes be conveniently found by multiplying the required roots of the numbers.)

(h) 0; 8; 125

(*Note.*—The log of 0 is $-\infty$ and the reciprocal of 0 is infinity, hence the geometric and harmonic means of a series containing 0 are each 0.)

8. (a) Find the geometric mean of the following distribution:

m	f	$\log m$
2	1	0.3010
4	5	0.6021
6	3	0.7782
8	1	0.9031

(b) Find the geometric mean of

$$m = 3, 4, 5, 6, 7, 8, 9, 10; \ f = 1, 14, 25, 27, 18, 9, 4, 2.$$

9. (a) Find the quadratic mean of the following deviations:

$-3;\quad -1;\quad 0;\quad 1;\quad 3$

(b) Find the quadratic mean of the deviations of the following numbers from their arithmetic mean:

2; 4; 6; 8; 10; 12; 14

10. (a) What are the medians of the series of numbers in the preceding exercise?
(b) What is the median of the following numbers:

11; 4; 9; 15; 6; 12; 8; 5; 10

(c) Find *AM* and *Md* of both scales of the double frequency table of Exercise 7, p. 31 (Chapter II).

11. I. Find the median and mode of each of the following distributions:

(a) *m*	*f*	(b) *m*	*f*	(c) *m*	*f*	(d) *m*	*f*	(e) *m*	*f*
2	1	20	1	2	4	1	1	10	2
4	4	40	5	4	7	2	4	12	5
6	6	60	3	6	5	3	5	14	6
8	4	80	1	8	3	4	3	16	4
10	1			10	1	5	2	18	2
						6	1	20	1

II. Find the arithmetic, geometric, and harmonic means of the preceding distributions.

12. Check graphically the calculations of Exercise 11—I.

13. Find the arithmetic means of the distributions of Exercises 1 to 6 inclusive, pp. 26–30 (Chapter II).

14. Find the medians and modes of the same distributions.

15. Find the arithmetic, geometric, and harmonic means of the three total distributions, Exercise 5, pp. 27–29 (Chapter II).

16. Find the *AM*, *GM*, *HM*, *Md*, and *Mo* of the distributions given in Exercise 1, p. 85, Exercise 4, p. 86, and Exercise 5, p. 86 (Chapter IV).

17. Given the following rates of increase per decade in the United States, 1790 to 1860, find the average rate of increase per decade; also the average rate of increase per year. Prove the answers.

Decade	Per cent increase	Decade	Per cent increase
1790–1800	35.1	1830–1840	32.7
1800–1810	36.4	1840–1850	35.9
1810–1820	33.1	1850–1860	35.6
1820–1830	33.5		

18. An interesting example of the use of the median is in the determination of the weekly median change in prices experienced by all the common stocks on the New York Stock Exchange. The quartiles may also be calculated from week to week to show the varying range of price changes (*cf.* page 73).

In the following problem the weekly data for the month of September, 1930, have been tabulated and the frequencies recorded. That is, in the weekly stock quotations

(per the *Annalist*) the weekly change (dollars per share) is tabulated for each common stock listed and traded in. In the tabulation here used (cf. Table 7), the interval is $\frac{1}{2}$, with the exception of the no-change class in which the theoretical interval is $\frac{1}{8}$. The theoretical fractional limits have been reduced to decimals for convenience. The theoretical limits of the first finite class is $-5\frac{1}{16}$ to $-4\frac{9}{16}$, or -5.0625 to -4.5625. Included in this class are the actual changes of -5, $-4\frac{7}{8}$, $-4\frac{3}{4}$, and $-4\frac{5}{8}$. The limits are set mid-way between two actual quotations.

TABLE 7

Tabulation of weekly changes in the prices of stocks

Theoretical limits		Actual changes per *Annalist* included within theoretical limits				Number of changes in each class during weeks ending September				
L_1	L_2					7 f	7 Σ_1*	13 f	20 f	27 f
−5.5625	and less	$-5\frac{1}{2}$	....	....	....	5	0	8	30	105
−5.0625	−4.5625	-5	$-4\frac{7}{8}$	$-4\frac{3}{4}$	$-4\frac{5}{8}$	1	5	1	12	20
−4.5625	−4.0625	$-4\frac{1}{2}$	$-4\frac{3}{8}$	$-4\frac{1}{4}$	$-4\frac{1}{8}$	1	6	2	11	21
−4.0625	−3.5625	-4	$-3\frac{7}{8}$	$-3\frac{3}{4}$	$-3\frac{5}{8}$	0	7	5	16	38
−3 5625	−3.0625	$-3\frac{1}{2}$	$-3\frac{3}{8}$	$-3\frac{1}{4}$	$-3\frac{1}{8}$	2	7	6	17	39
−3.0625	−2.5625	-3	$-2\frac{7}{8}$	$-2\frac{3}{4}$	$-2\frac{5}{8}$	3	9	11	28	62
−2.5625	−2.0625	$-2\frac{1}{2}$	$-2\frac{3}{8}$	$-2\frac{1}{4}$	$-2\frac{1}{8}$	3	12	6	40	69
−2.0625	−1.5625	-2	$-1\frac{7}{8}$	$-1\frac{3}{4}$	$-1\frac{5}{8}$	13	15	30	74	76
−1.5625	−1.0625	$-1\frac{1}{2}$	$-1\frac{3}{8}$	$-1\frac{1}{4}$	$-1\frac{1}{8}$	20	28	36	90	77
−1.0625	−0.5625	-1	$-\frac{7}{8}$	$-\frac{3}{4}$	$-\frac{5}{8}$	55	48	79	97	99
−0.5625	−0.0625	$-\frac{1}{2}$	$-\frac{3}{8}$	$-\frac{1}{4}$	$-\frac{1}{8}$	104	103	141	132	77
−0.0625	+0.0625	0	....	....	....	89	207	79	57	29
+0.0625	+0.5625	$+\frac{1}{8}$	$+\frac{1}{4}$	$+\frac{3}{8}$	$+\frac{1}{2}$	136	296	110	36	13
+0.5625	+1.0625	$+\frac{5}{8}$	$+\frac{3}{4}$	$+\frac{7}{8}$	$+1$	82	432	73	31	7
+1.0625	+1.5625	$+1\frac{1}{8}$	$+1\frac{1}{4}$	$+1\frac{3}{8}$	$+1\frac{1}{2}$	49	514	49	10	3
+1.5625	+2.0625	$+1\frac{5}{8}$	$+1\frac{3}{4}$	$+1\frac{7}{8}$	$+2$	41	563	26	7	1
+2.0625	+2.5625	$+2\frac{1}{8}$	$+2\frac{1}{4}$	$+2\frac{3}{8}$	$+2\frac{1}{2}$	20	604	8	2	0
+2.5625	+3.0625	$+2\frac{5}{8}$	$+2\frac{3}{4}$	$+2\frac{7}{8}$	$+3$	15	624	10	3	0
+3 0625	+3.5625	$+3\frac{1}{8}$	$+3\frac{1}{4}$	$+3\frac{3}{8}$	$+3\frac{1}{2}$	6	639	3	1	2
+3.5625	+4.0625	$+3\frac{5}{8}$	$+3\frac{3}{4}$	$+3\frac{7}{8}$	$+4$	14	645	6	0	0
+4.0625	+4.5625	$+4\frac{1}{8}$	$+4\frac{1}{4}$	$+4\frac{3}{8}$	$+4\frac{1}{2}$	2	659	4	2	1
+4.5625	+5.0625	$+4\frac{5}{8}$	$+4\frac{3}{4}$	$+4\frac{7}{8}$	$+5$	3	661	5	0	1
+5.0625	and up	$+5\frac{1}{8}$	....	....	....	11	664	9	2	3
						675	675	707	698	743

* A convenient short-cut method of obtaining the Σ_1 items is through the use of an adding machine on which the sub-totals are taken after each frequency is added; thus the Σ_1 and Σ_2 of each class are recorded on the tape. In the case of many machines these figures are in red. If greater speed is desired the frequencies at the extremes need not be sub-totaled when it is seen by inspection that the quartiles do not fall in those classes.

Applying the formulas for the finding of the median and quartiles the following results are obtained for the week ending September 7.

The quartile points are: $n/4 = 168.75$; $n/2 = 337.5$; $3n/4 = 506.25$.

$$Q_1 = L_1 + i(n/4 - \Sigma_1)/f = -\ 0.5625 + \tfrac{1}{2}(168.75 - 103)/104$$
$$= -\ 0.5625 + 0.3161 = -\ 0.2464$$

$$Q_2 = L_1 + i(n/2 - \Sigma_1)/f = +\ 0.0625 + \tfrac{1}{2}(337.5 - 296)/136$$
$$= +\ 0.0625 + 0.1526 = +\ 0.2151$$

$$Q_3 = L_1 + i(3n/4 - \Sigma_1)/f = +\ 0.5625 + \tfrac{1}{2}(506.25 - 432)/82$$
$$= +\ 0.5625 + 0.4527 = +\ 1.0152$$

Applying the same method, compute the medians and quartiles for the weeks ending September 13, 20, and 27.

These results may be cumulated into an index of price change by adding the medians, week after week. It is somewhat more difficult to obtain the quartile indexes.

19. The following tables were taken from "Statistics of Income for 1929," compiled by the Bureau of Internal Revenue, United States Treasury Department, p. 5. The tables show the distribution, by a limited number of net income classes, of the number of returns filed, the amount of net income, and tax reported as well as the cumulative totals and relative percentages.

(a) Compute the arithmetic mean of net income and tax returns. What do they signify?

(b) Compute the median and mode for (1) returns made, (2) net income, and (3) tax returns. In each case what do the median and mode signify?

(c) Plot the three distributions. Determine graphically the median and mode. Check with the computed results.

STATISTICS OF INCOME

Simple and cumulative distribution of individual returns for 1929, by net income classes, showing number of returns, net income, tax, and percentages

New income classes (Thousands of dollars)	Returns					
	Simple distribution		Cumulative distribution over class below		Cumulative distribution under class above	
	Number	Per cent	Number	Per cent	Number	Per cent
Under 1 (estimated)	126,172	3.12	4,044,327	100 00	126,172	3.12
1 under 2 (estimated)	903,082	22.33	3,918,155	96.88	1,029,254	25.45
2 under 3 (estimated)	810,347	20.04	3,015,073	74.55	1,839,601	45.49
3 under 5 (estimated)	1,172,655	29.00	2,204,726	54.51	3,012,256	74.49
5 under 10	658,039	16.27	1,032,071	25.51	3,670,295	90.76
10 under 25	271,454	6.71	374,032	9.24	3,941,749	97.47
25 under 50	63,689	1.57	102,578	2.53	4,005,438	99.04
50 under 100	24,073	0.60	38,889	0.96	4,029,511	99.64
100 under 150	6,376	0.16	14,816	0.36	4,035,887	99.80
150 under 300	5,310	0.13	8,440	0.20	4,041,197	99 93
300 under 500	1,641	0.04	3,130	0.07	4,042,838	99.97
500 under 1000	976	0.02	1,489	0.03	4,043,814	99.99
1000 and over	513	0.01	513	0.01	4,044,327	100 00
Total	4,044,327	100.00				

STATISTICS OF INCOME—*Continued*

Net income classes (Thousands of dollars)	Net income					
	Simple distribution		Cumulative distribution over class below		Cumulative distribution under class above	
	Amount	Per cent	Amount	Per cent	Amount	Per cent
Under 1 (estimated) .	$73,742,132	0.30	$24,800,735,564	100.00	$73,742,132	0.30
1 under 2 (estimated).	1,499,907,745	6.05	24,726,993,432	99.70	1,573,649,877	6.35
2 under 3 (estimated).	1,958,594,897	7.90	23,227,085,687	93.65	3,532,244,774	14.25
3 under 5 (estimated).	4,572,596,263	18.44	21,268,490,790	85.75	8,104,841,037	32 69
5 under 10..........	4,481,575,786	18.07	16,695,894,527	67.31	12,586,416,823	50 76
10 under 25..........	4,025,233,375	16.23	12,214,318,741	49.24	16,611,650,198	66 99
25 under 50..........	2,174,458,126	8.77	8,189,085,366	33.01	18,786,108,324	75.76
50 under 100........	1,646,476,000	6.64	6,014,627,240	24.24	20,432,584,324	82.40
100 under 150........	770,536,078	3.11	4,368,151,240	17.60	21,203,120,402	85 51
150 under 300........	1,087,409,737	4.38	3,597,615,162	14.49	22,290,530,139	89.89
300 under 500........	628,228,889	2.53	2,510,205,425	10.11	22,918,759,028	92.42
500 under 1000	669,877,752	2.70	1,881,976,536	7.58	23,588,636,780	95.12
1000 and over........	1,212,098,784	4.88	1,212,098,784	4.88	24,800,735,564	100.00
Total.............	$24,800,735,564	100.00				

Net income classes (Thousands of dollars)	Tax					
	Simple distribution		Cumulative distribution over class below		Cumulative distribution under class above	
	Amount	Per cent	Amount	Per cent	Amount	Per cent
Under 1 (estimated)..	$17,308	0.01	$1,001,938,147	100 00	$17,308	0 01
1 under 2 (estimated).	553,418	0.06	1,001,920,839	99.99	570,726	0.07
2 under 3 (estimated).	1,403,901	0.14	1,001,367,421	99.93	1,974,627	0 21
3 under 5 (estimated).	2,412,634	0.24	999,963,520	99.79	4,387,261	0.45
5 under 10..........	9,550,599	0.95	997,550,886	99 55	13,937,860	1 40
10 under 25..........	59,893,017	5.98	988,000,287	98.60	73,830,877	7.38
25 under 50..........	113,904,197	11.37	928,107,270	92.62	187,735,074	18.75
50 under 100.........	160,813,524	16.05	814,203,073	81.25	348,548,598	34.80
100 under 150........	99,559,757	9.94	653,389,549	65 20	448,108,355	44.74
150 under 300........	159,221,214	15.89	553,829,792	55.26	607,329,569	60.63
300 under 500	97,335,662	9.71	394,608,578	39.37	704,665,231	70.34
500 under 1000......	106,218,910	10.60	297,272,916	29.66	810,884,141	80.94
1000 and over........	191,054,006	19.06	191,054,006	19.06	1,001,938,147	100.00
Total....	$1,001,938,147	100.00				

ANSWERS

1. (a) 15
(b) 4
(c) 4
(d) 10
(e) $17\frac{1}{2}$
(f) 4
(g) −1
(h) 9
(i) 10
(j) 37
(k) 0.41
(l) 0.66

2. (a) $505\frac{1}{4}$
(b) 686,709
(c) 5.2
(d) 4.4

3. (a) $4.80
(b) 4.80

4. (a) $6.00
(b) $4.50

5. (a) $0.64
(b) 0.47
(c) 2.11
(d) 1.56

6. (a) 32.5
(b) 30.8
(c) 12.0

7.

	GM	*AM*	*HM*
(a)	6.32	7	5.714
(b)	94.57	$107\frac{1}{2}$	84.211
(c)	32	85	12
(d)	100	2222.2	4.5
(e)	6	$6\frac{1}{2}$	5.6
(f)	30	53.3	17.6
(g)	120	$244\frac{1}{2}$	49.8
(h)	0	44.3	0

8. (a) 4.52
(b) 5.83

9. (a) 2
(b) 4

10. (a) 0 and 8
(b) 9
(c) Value: *AM* 158.53, *Md* 152.73
Ratio: *AM* 48.087, *Md* 46.669

11.

	Md	*Mo*	*AM*	*GM*	*HM*
(a)	6	6	6	5.615	5.161
(b)	46	43.3	48	45.177	42.105
(c)	4.7	4.2	5	4.477	3.954
(d)	3.1	2.8	3.25	2.973	2.673
(e)	14	$13\frac{2}{3}$	14.2	13.966	13.735

12. For method, see Chart 8, p. 45. The check is theoretically exact.

13. (1) \$11.20; \$13.20
(2-a) 4
(2-b) 5.8
(3) 6
(4-a) 142
(4-b) 71.5

(5-A)	I. 147.64;	II. 140.25;	III. 199.37;	
	IV. 189.33;	V. 155.68;	VI. 193.33;	
	VII. 149.23;	VIII. 212.90;	IX. 144.8;	Σ166.60
(5-B)	155.36;	143.21;	192.92;	
	182.67;	148.65;	185.26;	
	143.08;	201.94;	139.20;	163.93
(5-C)	165.33;	157.53;	191.04;	
	182.67;	147.57;	184.21;	
	139.09;	193.75;	149.63;	167.18
(6)	100.42;	98.07;	103.59;	96.28
	95.34			

14.

	Mode	Median
(1-a)	$10\frac{2}{3}$	11
(1-b)	$12\frac{2}{3}$	13
(2-a)	3.45	3.7
(2-b)	4.07	5.75
(3)	5.68	5.87
(4-a)	136.67	140
(4-b)	70	71

(5-a)

Mode			
I. 135.71;	II. 123.04;	III. 212.00;	
IV. 200.00	V. 162.00;	VI. 180.00;	
VII. 117.50;	VIII. 230.00;	IX. 142.00;	Σ162.14
Median			
140.00;	137.33;	202.67;	
193.33;	160.00;	190.00;	
125.45;	215.00;	142.86;	163.44

(5-b)	Mode			
	144.67;	136.67;	210.00;	
	160.00;	146.00;	182.31;	
	101.43;	176.67;	138.57;	145.79
	Median			
	147.65;	141.36;	194.67;	
	173.33;	147.50;	185.00;	
	121.67;	192.50;	137.14;	159.10
(5-c)	Mode			
	171.82;	127.50;	179.09;	
	162.00;	140.67;	180.00;	
	101.67;	173.33;	140.00;	172.78
	Median			
	163.85;	156.43;	187.06;	
	167.50;	143.57;	182.00;	
	118.57;	184.29;	145.00;	165.57

15.

	AM	*GM*	*HM*
(A)	16.660	15.919	15.133
(B)	16.393	15.653	14.786
(C)	16.718	16.075	15.398

16.

(a)	*AM*	*GM*	*HM*	*Md*	*Mo*
a.	3.85	3.65	3.448	3.833	3.833
b.	8.30	7.92	7.500	8.333	8.333
c.	2.92	2.708	2.467	2.900	2.900
d.	6.17	5.739	5.199	6.200	6.200
(b)					
a.	21.000	18.392	15.528	20.000	17.5
b.	21.600	18.331	14.717	21.000	20.0
c.	5.667	5.395	5.108	5.600	5.5
d.	5.333	5.018	4.645	5.400	5.5
e.	10.750	10.589	10.427	10.667	10.5
f.	10.000	8.951	7.908	9.428	8.4
g.	26.000	23.646	21.169	25.000	23.0
(c)					
a.	6.9	6.206	5.451	6.600	5.667
b.	8.6	8.020	7.428	8.286	8.000
c.	8.6	8.086	7.560	8.222	7.333
d.	7.6	6.817	5.929	7.333	6.333

17. 34.6% per decade
3.01% per year

CHAPTER IV

DISPERSION

In the preceding chapter it was observed that an average serves the purpose of summarizing in a single item the magnitudes appearing in a series or tabulation of data. Whether it is of the strict mathematical form or not, it provides a measure of the central tendency of the items and gives a picture of them from this point of view. But as was previously pointed out, an average may be very misleading. In many cases the nature of the distribution or the degree of variability may be more important than the average itself. If a thousand shots are fired at a target and five hundred of them strike on the right and five hundred on the left, their average may be the same as that of a thousand shots striking the bull's-eye, but their variability discounts the value of the average. This illustration is extreme because a tabulation, as a rule, shows a maximum concentration at or close to the average, hence the average generally has in itself a high degree of significance. But it has much greater significance when considered with respect to a charted or calculated description of the distribution, and a mathematical measurement of the dispersion. This chapter is devoted to a consideration of the methods employed in measuring the dispersion or scatter of the data from the point of central tendency; that is, it studies the measurement of variability. It deals chiefly with unclassified and tabulated groups of data, but as we shall see later, the measures here discussed may also be applied to dispersions about a trend line as well as about a point of central tendency.

The range of variability.—Probably the first suggestion arising in the mind of the student in connection with the problem of variability is that the mere range—that is, a statement of the largest and smallest items and the difference between them—may be chosen as a suitable measure. This range is very evident in a chart of a distribution, and superficially appears to be the most natural measure of dispersion. But it is not a dependable measure of the degree of variability except when the distribution is markedly regular, since there are often a few scattering items at the extremes which carry the range significantly

beyond that which is normal to the entire distribution. Or, again, the extremes may be cut off abruptly. Hence, as a rule, the range itself is not of much value except as a very rough measure. Any adequate measure of variability should not be dependent simply upon the position of two or three erratic items, but should take into account practically all of the items.

The average deviation (*AD*).—The simplest measure of dispersion is the average deviation. This is merely a common average of the deviations of the items (m) from the central average or other point of origin (R), the deviations being taken as positive, as is illustrated by both untabulated and tabulated data in Example 15. The average deviation is usually taken from either the arithmetic mean or the median as a measure of central tendency. If the arithmetic mean is chosen as the origin of the deviations, the sum of the positive deviations equals the sum of the negative. On the other hand, if the median is taken as the origin, the sum of the deviations considered irrespective of their signs is a minimum, in the sense, at least, that no other origin will give a smaller sum.* That is,

$$\Sigma(m - AM) = 0$$

$$\Sigma|m - Md| = \text{a minimum}$$

if, when n is even, other points in the median space are neglected. Other origins, such as the mode, may sometimes be employed, but they have little significance, except in special cases where the measurement of skewness is involved.

Example 15.—The average deviation (AD). $AD = \Sigma'd'/n$ where $'d' = |m - R|$, i.e., the absolute deviations of m's from the origin. The arithmetic mean is taken as origin (R) in parts (a) and (b), while the median is taken as origin in (c).

* In tabulated data, the average deviation computed from the mean and the median, respectively, will not always indicate that the median as origin gives a minimum. The reason for this is that the median is interpolated on the assumption that the items are distributed regularly through the median class, whereas the average deviation is found on the assumption that they are concentrated at the class mark. However, the correction c_1 (cf. Example 20, p. 80) makes the former assumption uniform for both calculations. Thus in Example 3a (Laboratory Exercises at the close of this chapter), the mean is 12 and the average deviation from it is 2.24, while the median is 11.741, and the average deviation from it is 2.292. But if the correction c_1 is added, the average deviations become 2.375 and 2.366, respectively. However, it is hardly practicable to use the correction c_1 unless c_2 is also used. These corrections are seldom required in practice.

(a) Untabulated data, *AM* as *R*.

m	$d = (m - AM)$
8	1
6	−1
10	3
7	0
4	−3
$AM = 7$	$5\overline{)8} = \Sigma' d'$
	$1.6 = AD$

AD expressed as a coefficient or percentage is

$$\text{Coef. } AD = AD/R = 1.6/7 = 0.23 = 23\%$$

(b) Tabulated data, *AM* as *R*.

m	f	fm	d	fd
3	20	60	−2.6	−52
5	40	200	−0.6	−24
7	30	210	1.4	42
9	10	90	3.4	34
	100	$\overline{)560}$		$100\overline{)152} = \Sigma' d'$
		$5\,6 = AM$		$1.52 = AD$

AD expressed as a coefficient or percentage is

$$\text{Coef. } AD = AD/R = 1.52/5.6 = 0.27 = 27\%$$

(c) Tabulated data, *Md* as *R*.

	L_1	L_2	m	f	Σ_1	Σ_2	$d = (m - Md)$	fd
	2	4	3	20	0	20	−2.5	−50
(*MdC*)	4	6	5	40	20	60	−0.5	−20
	6	8	7	30	60	90	1.5	45
	8	10	9	10	90	100	3.5	35
	$i = 2$			$n = 100$				$100\overline{)150} = \Sigma' d'$
				$n/2 = 50$ *Md* space				$1.5 = AD$

$$Md = L_1 + i(n/2 - \Sigma_1)/f \text{ in median class } (MdC)$$
$$Md = 4 + 2(50 - 20)/40 = 5.5$$
$$\text{Coef. } AD = 1.5/5.5 = 0.27 = 27\%$$

When the median is taken as the origin of the deviations, the process for untabulated data may be abbreviated by simply subtracting the sum of all items (S_1) smaller than the median, from the sum of all items (S_2) larger than the median, and dividing by the total number of items; i.e., $AD = (S_2 - S_1) \div n$. If the median is one of the

items, it is not included in the sums, but is counted in n. For example, the average deviation (from $Md = 7$) of 4, 6, 7, 8, 10 is:

$$AD = [(10 + 8) - (6 + 4)] \div 5 = 8/5 = 1.6$$

This method may be adapted to deviations from any given origin by including the origin in the sum having the fewer items enough times to equalize the number of items in each sum. The result is divided by the actual number of items (n) as before. For example, in the series just cited, the average deviation from the arbitrary origin, 9, is:

$$AD = [(10 + \mathit{9} + \mathit{9} + \mathit{9}) - (8 + 7 + 6 + 4)] \div 5 = 12/5 = 2.4$$

This adaptation may be applied to the average deviation of tabulated data from any origin by comparing the fm's above and below the origin, equalizing the frequencies above and below by inserting the origin as in the preceding illustration. For example, in the illustration of the average deviation of tabulated data above (Example 15, b), we might have taken:

$$AD = [(90 + 210 + \mathit{20} \times \mathit{5.6}) - (200 + 60)] \div 100 = 1.52$$

The term 20×5.6 inserts the origin ($AM = 5.6$) enough times to equalize the frequencies in the two groups (40 in the larger magnitudes, and 60 in the smaller). The process is illustrated in detail in Example 16. A graphic method is described later in connection with a discussion of percentiles (cf. p. 74).

Example 16.—The short-cut method of computing the average deviation, taking the arithmetic mean as origin (R).

(a) By class marks and frequencies. The mean is calculated directly as $\Sigma mf/n$ and is inserted serially in the m column (if identical with an m, either above or below it). In the f column opposite the mean the absolute difference between the frequencies preceding and following is inserted. A horizontal line is then drawn dividing the frequency column, as thus corrected, into two equal totals. The mean times its inserted frequency is entered as a correction in the mf column. Then Σmf above the line is subtracted from Σmf below the line ($\Sigma' d' = S_2 - S_1$). The horizontal line may be omitted if the inserted frequency (f_i) is given the sign of the difference of the total preceding (Σf_p) less the total following frequencies ($f_i = 2\Sigma f_p - n$).

(b) By step deviation. The same principle may be applied to any simplified deviation scale, preferably from an assumed origin near the mean, and expressed in class interval units. The mean may be obtained from Σfd as in short-cut averaging. The mean of the d column is inserted and $\Sigma' d'$ obtained

as before. Since $\Sigma' d'$ is in class interval units, it is multiplied by the class interval ($i = 2$).

(a)

	m	f	mf	
	3	2	6	
	5	4	20	$26 = S_1$
AM (R)	5.6	(2)	(11.2)	
	7	3	21	
	9	1	9	$41.2 = S_2$
		10	)56	
		AM =	5.6	

$\Sigma' d' = S_2 - S_1 = 15.2$
$AD = \Sigma' d'/n = 15.2/10 = 1.52$

(b)

	m	f	d/i	fd/i	
	3	2	-1	-2	
$M_a =$	5	4	0	0	$-2 = S_1$
		(2)	$c/i = 0.3$	(0.6)	
	7	3	1	3	
	9	1	2	2	$5.6 = S_2$
				10)3	
			$c/i = 0.3$		

$\Sigma' d'/i = S_2 - S_1 = [5.6 - (-2)] = 7.6$
$\Sigma' d' = 7.6i = 15.2$
$AD = \Sigma' d'/n = 15.2/10 = 1.52$
$AM = M_a + c = 5 + 0.3 \times 2 = 5.6$

Several variations of the method of computing the average deviation are possible, as will be discovered if considerable machine calculation is done. For example, it is often convenient to take as deviations the steps: 0, 1, 2, 3, etc., and thus avoid negative signs. This procedure is equivalent to assuming the smallest class mark as the average (A_a), and requires the adjustments relative to the class interval indicated in Example 16 (b). A convenient method to use as a check, when the mean is the origin, is the computation of the actual deviations times the frequencies on one side of the mean only. This will give half the total deviations ($\Sigma' d'/2$).

It is sometimes desirable to compute the average deviation separately on each side of the origin, in order to give some idea of the skewness. When this is done, the most suitable origin may prove to be the mode. The short-cut methods may be adapted to the required computation, but since the process does not furnish a standard measure of skewness, it need not be considered in detail here.

The standard deviation (SD or σ).—As a measure of dispersion in the field of the social sciences, the average deviation is usually quite satisfactory, particularly if the frequencies depart considerably from the normal or logarithmic normal types. But for large and highly regular distributions the so-called standard deviation is mathematically more desirable; in fact, in many advanced statistical calculations

involving complex formulas, it is essential. The standard deviation may be defined as a quadratic mean of the deviations; that is, it is the square root of the average squared deviation ($\sigma^2 = \Sigma d^2/n$). It is computed from the arithmetic mean as origin (AM as R); if another origin is used, the measure of deviation thus obtained is more precisely designated by the more general term, root-mean squared deviation. It will be seen that by the process of squaring the deviations, the negative signs disappear, and the final square root has a plus or minus value which is, however, usually written as positive. The standard deviation as thus computed from the mean as origin is a minimum, and it is seldom that any other origin is appropriate. It should be added that in many problems in which the standard deviation enters, the square of the standard deviation ($\Sigma d^2/n$) is employed. This measure is called the variance.

In the case of an untabulated series, the deviations are taken by subtracting the origin from each item ($d = m - AM$). The deviations are then squared and averaged ($\sigma^2 = \Sigma d^2/n$), and the square root of this average is taken. But since the deviations thus found may involve several decimal places and may thus be inconvenient to square, it is advisable with small numbers to make use of an algebraic equivalent by which the items themselves are squared and averaged, and the resulting average is then decreased by the square of the arithmetic mean; that is, $\sigma^2 = \Sigma d^2/n = \Sigma m^2/n - AM^2$.

Example 17.—The standard deviation. $\sigma^2 = \Sigma d^2/n = \Sigma m^2/n - AM^2$, where the deviations are taken from the mean of the items (m); that is, $d = m - AM$. For untabulated data (a), and tabulated data (b). A short-cut method is given later. It is evident that m^2f may be obtained as m times mf; or, to check, as m^2 times f. The total of this column is Σm^2, which, divided by n and diminished by AM^2, gives σ^2. The tabulated form (b) is identical in principle with (a).

(a) Untabulated data — (or) Without d and d^2

m	d	d^2	m^2
6	−1	1	36
10	3	9	100
7	0	0	49
4	−3	9	16
8	1	1	64
5)35	0	5)20	5)265
$AM = 7$		$\sigma^2 = 4$	53
		$\sigma = 2$	$AM^2 = 49$
			$\sigma^2 = 4$
			$\sigma = 2$

Coef. $\sigma = 2/7 = 0.29$

(b) Tabulated data (or) Without d, d^2, and fd^2.

m	f	mf	d	d^2	fd^2	m^2f
3	20	60	−2.6	6.76	135.2	180
5	40	200	−0.6	0.36	14.4	1,000
7	30	210	1.4	1.96	58.8	1,470
9	10	90	3.4	11.56	115.6	810
	100	100)560			100)324.0 $= \Sigma d^2$	100)3,460 $= \Sigma m^2$
		$AM = 5.6$			$\sigma^2 = 3.24$	$= 34.60$
					$\sigma = 1.8$	$AM^2 = 31\ 36$
						$\sigma^2 = 3\ 24$
						$\sigma = 1\ 8$

Coef. $\sigma = 1.8 \div 5.6 = 0.32$

Short-cut method of standard deviation.—The most common method of computing the standard deviation employs an assumed mean and takes the deviations from it. The standard deviation is then found on the basis of these deviations and the frequencies. Since changing from the actual items to any set of deviations from an arbitrary origin does not affect the scatter of the data, the calculation may be carried out by an adaptation of the method expressed by the formula used in Example 17:

$$\sigma^2 = \Sigma m^2 \div n - AM^2$$

When applied to the d column, the average of which is c, this formula becomes:

$$\sigma^2 = \Sigma d^2 \div n - c^2$$

If the assumed average happens to be the correct average, then the d column gives the true deviations on which the standard deviation is directly based, and $c = 0$. The short-cut method may be applied to either untabulated or tabulated data as illustrated in Example 18.

Example 18.—Standard deviation; short-cut method. $\sigma^2 = \Sigma d^2 \div n - c^2$, where the deviations ($d = m - M_a$) are taken from any convenient assumed origin, preferably near the average, and c is the average of the d's thus obtained. In tabulated data, f is assumed with Σ. The fd^2 column may be found as d times fd; or, to check, as d^2 times f. It may also be noted that the standard deviation of tabulated data may be found by expressing the d column in units of class intervals, as: -1; 0; 1; 2; or 0; 1; 2; 3. When this is done, the computation, including c and σ, is expressed in class interval units (c/i and σ/i); hence both are multiplied by i in finding $AM = M_a + c$, and σ.

(a) Untabulated data.

m	d	d^2
6	-2	4
10	2	4
7	-1	1
4	-4	16
$M_a = 8$	0	0
	$5)\overline{-5}$	$5)\overline{25}$
	$c = -1$	5
	$M_a = 8$	$c^2 = 1$
	$AM = 7$	
		$\sigma^2 = 4$
		$\sigma = 2$

(b) Tabulated data.

m	f	d	fd	fd^2
3	20	-2	-40	80
$M_a = 5$	40	0	0	0
7	30	2	60	120
9	10	4	40	160
	$n = 100$		$100)\overline{60}$	$100)\overline{360} = \Sigma d^2$
			$c = 0.6$	3.6
				$c^2 = 0.36$
				$\sigma^2 = 3.24$

$AM = 5 + 0.6 = 5.6$ $\qquad$ $\sigma = 1.8$ Coef. $\sigma = 1.8/5.6 = 0.32$

The standard deviation, since it employs the second moment * $(\Sigma d^2/n)$ about the mean, puts more emphasis on the larger deviations, and gives a larger result than the average deviation (normally $AD = 0.7979\ \sigma$). Both the average deviation and the standard deviation when computed from tabulated data tend to be a little larger than they would be if calculated from the original data before tabulation, since in each class the real average is usually not the class mark, but a point slightly closer to the mode, as determined by the tendency of the items in each class to cluster more densely on the modal side. Corrections for offsetting this tendency are discussed later, but they are not usually necessary in practice. The measurement of dispersion in tabulated data may also be distorted a little by marked irregularities in the data. Such distortions may perhaps best be avoided by the use of another measure, the so-called quartile deviation.

* The "first moment" is the algebraic average of the deviations of a series of items about an arbitrary origin or about the mean of the series. The second moment is the average of the squared deviations; the third moment the average of the cubed deviations, etc.

Quartile deviation (QD) and percentiles (P).—Just as the median (the 50th percentile or P_{50}) was interpolated in untabulated or tabulated data, so points may be located on the magnitude scale at any other percentiles, and the spread measured by reference to these points. The points most used for this purpose are the quartiles. The first quartile ($Q_1 = P_{25}$) is described as the actual or interpolated point below which 25% of the items fall, and the third quartile ($Q_3 = P_{75}$) similarly is the point below which 75% of the items fall. The second quartile ($Q_2 = P_{50}$) is identical with the median. One-half of the distance from the first to the third quartile is taken as the measure of quartile deviation ($QD = \overline{Q_3 - Q_1} \div 2$). In the normal symmetrical curve, this measure is 0.67449 σ, and as applied to sampling errors is known as the probable error. It measures the distances, above and below the average (from $AM - 0.6745\,\sigma$ to $AM + 0.6745\,\sigma$), which include half the items, or frequencies. The process of calculation is illustrated in Example 19. A method of measuring skewness by means of the quartiles is included in the example.

The 10–90 percentile range ($P_{90} - P_{10}$) is also used as a measure of dispersion. This measure includes four-fifths of the distribution, excluding only the erratic extreme items. Any percentile may be located or interpolated as is illustrated below for the quartiles. To illustrate, in Example 19, the tenth (P_{10}) and ninetieth (P_{90}) percentiles are:

$$P_{10} = L_1 + i(n/10 - \Sigma_1) \div f = 2 + 2(1 - 0) \div 2 = 3$$

$$P_{90} = L_1 + i(9n/10 - \Sigma_1) \div f = 8 \text{ (}9n/10\text{ falls on }\Sigma_1\text{; hence }P_{90} = L_1\text{)}$$

The specific notation in each formula refers to the class in which $n \div 10$ and $9n \div 10$ are located, as determined by the cumulatives. The 10–90 percentile range is therefore,

$$10\text{–}90\ P \text{ range} = P_{90} - P_{10} = 8 - 3 = 5$$

This obviously means that 80% of the workers receive wages between \$3 and \$8. If an abstract measure of the range is required, the 10–90 percentile range may be expressed in units of the median or other type $(P_{90} - P_{10}) \div Md$.

Example 19.—Quartile interpolation and deviation, $Q_1 = L_1 + i(n/4 - \Sigma_1) \div f$; $Q_3 = L_1 + i(3n/4 - \Sigma_1) \div f$; $QD = (Q_3 - Q_1) \div 2$ and coef. $QD = (Q_3 - Q_1) \div (Q_3 + Q_1)$, and skewness ($Sk$) may be measured by the formula: $(Q_3 + Q_1 - 2Q_2 \div (Q_3 - Q_1)$. The quartile class is located by the serial position of $n/4$ and $3n/4$ in the cumulatives. The notation of the formulas refers to the classes thus determined.

Classes		Frequency	Cumulatives		
L_1	L_2	f	Σ_1	Σ_2	
2	4	2	0	2	
4	6	4	2	6	Q_1 class
6	8	3	6	9	Q_3 class
8	10	1	9	10	
$i = 2$		$n = 10$			

$n/4 = 2.5$; $3n/4 = 7.5$.
$Q_1 = L_1 + i(n/4 - \Sigma_1) \div f = 4 + 2(2.5 - 2) \div 4 = 4.25.$
$Q_3 = L_1 + i(3n/4 - \Sigma_1) \div f = 6 + 2(7.5 - 6) \div 3 = 7.$
$QD = (Q_3 - Q_1) \div 2 = (7 - 4.25) \div 2 = 1.375.$

The quartile deviation, $(Q_3 - Q_1) \div 2$, expressed as a percentage of the mid-quartile point, $(Q_3 + Q_1) \div 2$, is the quartile coefficient of dispersion: Coef. $QD = (Q_3 - Q_1) \div (Q_3 + Q_1) = 2.75 \div 11.25 = 0.24$ or 24%.

The skewness (Sk) of the above distribution may be measured thus: $Sk = (Q_3 + Q_1 - 2Q_2) \div (Q_3 - Q_1) = (7 + 4.25 - 2 \times 5.5) \div (7 - 4.25) = 0.091 = 9.1\%$.

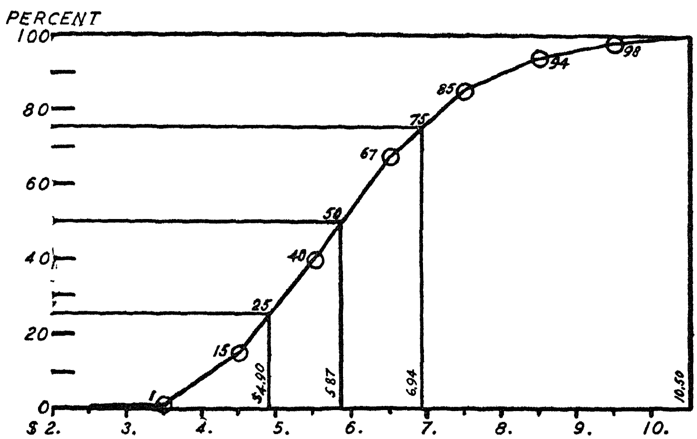


CHART 10

Cumulative chart of the distribution $m = 3, 4, 5, 6, 7, 8, 9, 10$ and $f = 1, 14, 25, 27, 18, 9, 4, 2$, with the interpolation of quartiles. The summations (small circles) are plotted at the class limits (Σ_2 at L_2); the quartile points, 25%, 50%, and 75%, are located on the vertical scale, horizontal lines are drawn to the cumulative curve, and perpendiculars are dropped from the points of intersection to the magnitude scale (X), thus locating the quartile magnitudes. The interpolation is here on a straight line, in conformity with the usual calculations (cf. Example 19). A more exact interpolation, which is equivalent to rounding the cumulative line on the Chart, is shown in Chart 10a.

Graphic interpolation of percentiles.—Interpolation of the quartiles or of any percentile may easily be accomplished graphically by drawing the cumulative chart, which is preferably expressed as percentages of n (cf. Chart 10). Since the vertical scale represents an ordinal array of the workers, as in a bar chart (the number magni-

fied until the bars become mere lines), the wage corresponding to any given percentile may be found by coordinate points. For example, as has already been suggested, the median wage may be found by drawing a horizontal line from 50 (vertical scale) to the cumulative curve, and dropping a perpendicular from the point of intersection to the base line. The median may be read on the horizontal scale at the foot of

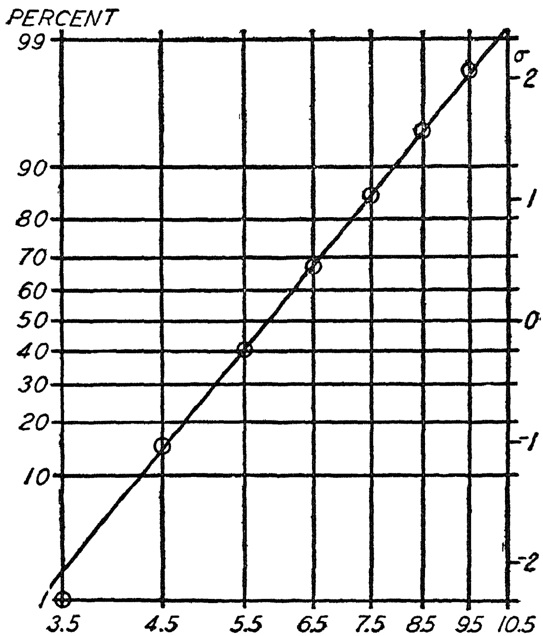


CHART 10a

Cumulative chart of the distribution $m = 3, 4, 5, 6, 7, 8, 9, 10$, and $f = 1, 14, 25, 27, 18, 9, 4, 2$ (cf. Chart 10) plotted on probability paper to a ratio horizontal scale. If probability paper with an arithmetic horizontal scale is used, the logs of L_2 should be plotted. The vertical scale is so arranged as to throw a normal distribution plotted against an arithmetic horizontal scale into a straight line; and a logarithmic normal distribution plotted against a logarithmic horizontal scale will similarly give a straight line. The chart therefore furnishes a test of the type of distribution, but considerable variation from a straight line outside of $+1\sigma$ and -1σ may be disregarded. The chart may be used for interpolating the quartiles and for estimating the frequencies of the normal curve fitted to the data.

this perpendicular. In the same way the other two quartiles, or any percentile, may be interpolated.

A similar interpolation of percentiles may be obtained more accurately by the use of probability paper,* which may be obtained from

* This graphic paper indicates departures from normality of distribution, on either the arithmetic or logarithmic magnitude scale, by departures of the cumulative curve from a straight line. Departures at the lower and upper extremes, however, are relatively unimportant, since the scale exaggerates them. Sometimes a logarith-

publishers in either the arithmetic or logarithmic form. The latter form is represented in Chart 10*a*. If arithmetic probability paper is used, the logarithms of the class limits and marks may be plotted, thus giving the same effect as the ratio scale. The arithmetic scale is used for normal distributions, and the logarithmic scale for logarithmic distributions. The cumulative curve expressed in percentages is plotted (Σ_2, against L_2 or its log), and the interpolation may be made as before. If the skewness is not logarithmic in form, the position of the curve on the chart may be experimentally shifted to the left or right along the logarithmic magnitude scale until the position is found at which the curve most closely approximates a straight line. In Chart 10*a* the cumulative points are so nearly in a straight line that a single line is drawn, but ordinarily the curve is drawn as a broken line; that is, separate straight lines are drawn from one point to the next as in the usual cumulative chart. The interpolation by probability paper is equivalent to smoothing each frequency from a rectangular form to a curve of the normal type, hence it is theoretically more accurate. The calculation may also be performed mathematically by the use of tables.

Summary.—Methods of measuring distributions for (a) point of central tendency, or type; and (b) degree of scatter, or dispersion, about the type taken as origin (R); assuming n items: $m_1 m_2 \ldots m_n$; and $d = m - R$.

(a) Type	(b) Dispersion
1. Arithmetic mean (AM) $AM = \Sigma m/n$	Standard deviation (σ) $\sigma^2 = \Sigma d^2/n = \Sigma m^2/n - AM^2$
2. Arithmetic mean (AM) $AM = \Sigma m/n$	Average deviation (AD) $AD = \Sigma' d'/n$
3. Median (Md) $Md = L_1 + i(n/2 - \Sigma_1)/f$	Average deviation (AD) $AD = \Sigma' d'/n$
4. Mid-quartile measure (MQ) $MQ = (Q_3 + Q_1)/2$	Quartile deviation (QD) $QD = (Q_3 - Q_1)/2$

mic distribution will tend to form a curve rather than a straight line. In such a case the quartiles may be tentatively read and a correction (c) obtained by the formula,

$$c = (Q_2^2 - Q_1Q_3) \div (Q_1 + Q_3 - 2Q_2)$$

If this correction is added to each of the class limits, and the cumulative curve again plotted, interpolations of percentiles may be obtained more accurately. Probability paper may also be used to fit a normal curve to data by drawing a straight line through the first and third quartile points, and reading the summations on this line at the ordinates of the class limits. The successive differences (first differences) of these summations are the required frequencies of the normal curve.

Other measures or combinations of measures are sometimes used, but the foregoing are the most important.

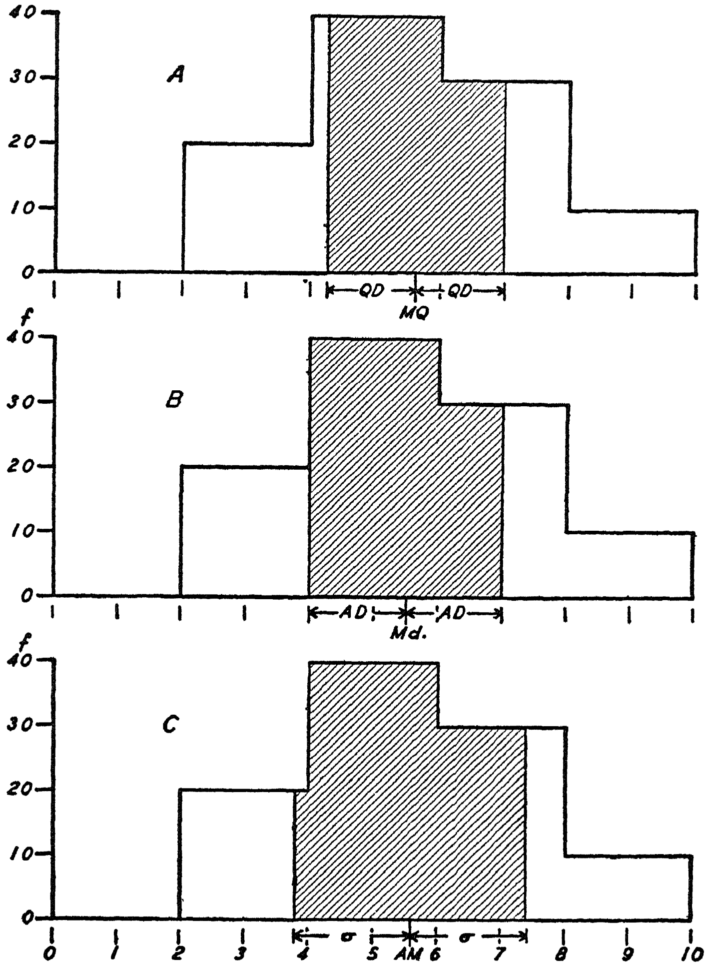


CHART 11

A comparison of quartile deviation, average deviation, and standard deviation for the distribution $m = 3, 5, 7, 9$; $f = 20, 40, 30, 10$. The points of origin, respectively, are the mid-quartile measure (MQ), the median (Md), and the arithmetic mean (AM). As Figure A indicates, the area included within one unit of quartile deviation above and below the origin is one-half the area of the total distribution. In normal distributions,

$$QD = 0.67449\sigma \quad \text{and} \quad AD = 0.79788\sigma.$$

The coefficient of deviation is obtained by dividing the dispersion by the types, as σ/AM, AD/AM, AD/Md; and $(Q_3 - Q_1)/(Q_3 + Q_1)$. The first combination of measures (AM and σ) is best adapted to

complete and regular data, particularly where further calculations of a complex nature are to be undertaken. It is sensitive to extreme items. With logarithmic distributions log m may be substituted for m, giving log G and log σ_r, the antilogs of which are the geometric mean and the standard deviation ratio. The last combination of measures (MQ and QD) is best adapted to inadequate and irregular data, being very insensitive to extreme items. The second (AM and AD) is perhaps best adapted to ordinary statistical work in the social sciences.

The standard deviation is a minimum about the arithmetic mean. The average deviation is a minimum about the median (with certain qualifications previously noted). The sum of the deviations is equal on both sides of the arithmetic mean. In a normal distribution, $AD = 0.7979\sigma$ and $QD = 0.6745\sigma$. In a normal logarithmic distribution, the geometric mean and the median are identical.

SUPPLEMENTARY METHODS

The methods of measuring dispersion which have already been discussed are sufficient for most practical purposes in the field of social statistics. It hardly needs to be pointed out that these methods are, broadly speaking, approximations only. As was previously suggested, it is inherent in the very nature of tabulations that this should be the case, since the frequency table groups the items by classes, and treats them as if on the average they had the magnitude of their respective class marks. A little experimentation will easily demonstrate that this is not precisely the case. As a rule there is a marked tendency for items to cluster toward the mode, so that a frequency distribution, with the minor irregularities smoothed out, is not as precisely represented by the usual rectangular chart as by one in which the frequencies slope toward the mode in somewhat the form suggested in Chart 12, which is a representation of a simple frequency distribution, similar to that previously used for purposes of illustration. The slope of each frequency toward the mode in this chart indicates roughly the probability that the items are more numerous on the side toward the mode, and less numerous on the outer side. An inspection of such a chart will make it clear that the class marks would represent the probability of distribution more accurately if they were moved a little toward the mode with a resulting decrease in the measure of dispersion. It is, of course, possible to measure approximately the required degree of such a readjustment. It is obvious, however, that if the classes are numerous and the class interval is small, the readjustment will be relatively unimportant.

Another minor inaccuracy in the calculation of measures of disper-

sion arises out of the class in which the arithmetic mean or other origin falls. If the origin happens to be identical, or nearly identical, with the class mark of this central class, then by the ordinary methods of computation scarcely any deviations will be recorded in this class, although it is obvious that the items within the class on each side of the origin might contribute a significant amount to the deviations. As the origin approaches a class limit, however, this error will diminish. Allowance can be made for this error by subdividing the central frequency

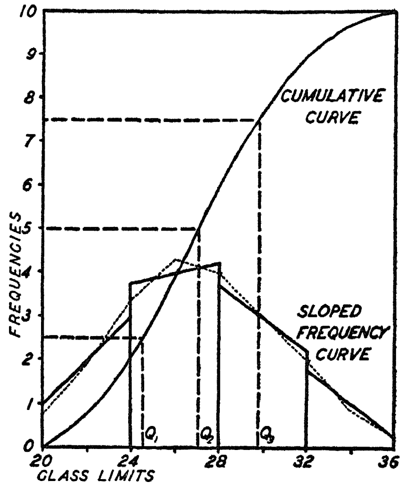


CHART 12

A sloped frequency curve. The distribution plotted is: m = 22, 26, 30, 34; f = 2, 4, 3, 1. The frequencies, instead of being plotted as rectangles, are made to slope toward the mode by giving to each the slope of the line joining the centers of the two adjacent frequencies. The dotted line carries the process a little farther by subdividing each frequency into two parts at the class mark. The cumulative curve is plotted in conformity with the sloped frequency curve. For details of the method, see the *Journal of the American Statistical Association*, December, 1929, p. 354.

into two parts at the point of origin, thus forming two sub-classes separately.

Another inaccuracy which has been given considerable attention by mathematicians arises out of the nature of sampling, particularly with reference to the second moment, or standard deviation. If the number of items in the sample is comparatively small, the standard deviation, σ_m, may be too small. It has been estimated that the error arising from this source is, as a rule, eliminated by dividing the variance by $n - 1$ instead of by n, i.e., $\sigma_m^2 = \sigma^2 \div (n - 1)$. In dealing with two or three constants, $n - 2$ and $n - 3$, respectively, would be employed as divisors. It will be seen that this correc-

tion becomes less important as the number of items increases. It is based upon the assumption of normal distributions, and since the tabulations generally met with in economic and social data are somewhat irregular logarithmic distributions, the correction is not employed in this book.

Corrections for *AD*.—In calculating an average deviation, corrections for the central class (c_1) and for the slope of the frequency (c_2) may be made by the use of special formulas. The derivation of these formulas has been described in the *Journal of the American Statistical Association* ("Analysis of Frequency Distributions," December, 1929, pp. 349–366). The formulas are given in Example 20 and are applied to a simple frequency distribution that has previously been used in illustrative examples. The result gives a corrected average deviation of 1.45 instead of 1.52 as obtained by the usual uncorrected method. Since the data referred to had a small number of classes and a small number of items, the correction is relatively large, but even in this case it is not of great significance.

Example 20.—The corrected average deviation, $AD_c = (\Sigma'd' + c_1 - c_2)/n$, where c_1 and c_2 are defined as given below and where d is the uncorrected deviation of the items from the mean taken as origin (R). The sum of the deviations is first computed by the usual method and is then corrected as indicated. The notation refers to the origin class, RC; $'v'$ is the absolute deviation of the class mark m, from the origin, R; D is the frequency slope at the origin (the frequency following the origin less the frequency preceding the origin), in this case $D = 3 - 4 = -1$; and f_a and f_z are the first and the last frequencies, respectively. If R is a class limit, either of the adjacent classes may be considered the origin class, and $c_1 = 0$.

Class		m	f	mf	
2– 4		3	2	6	
4– 6	RC	5	4	20	$26 = S_1$
	AM (R)	5.6	(2)	(11.2)	
6– 8		7	3	21	
8–10		9	1	9	$41.2 = S_2$
			10	)56	
$i = 2$				$AM = 5.6$	

$\Sigma'd' = S_2 - S_1 = 15.2$

$$c_1 = f(0.5i - 'v')^2 \div i = 4(0.5 \times 2 - 0.6)^2 \div 2 = 0.32.$$

$$c_2 = [fi - D(m - R)] \div 6 - i(f_a + f_z) \div 24$$

$$= [4 \times 2 + 1(5 - 5.6)] \div 6 - 2(2 + 1) \div 24 = 0.98.$$

$$AD \text{ (corrected)} = (\Sigma'd' + c_1 - c_2) \div n = (15.2 + 0.32 - 0.98) \div 10 = 1.45.$$

Corrections for *SD*, σ.—Corrections for class of origin and for slope, such as were made to the average deviation, assume a different form in

the case of the standard deviation, both corrections being combined in the formula for the total squared-deviations corrected (Σd_c^2), as follows:

$$\Sigma d_c^2 = \Sigma d^2 - i(ni + f_a d_p - f_z d_f)/12$$

where f_a and f_z are the first and last frequencies, respectively; and d_p and d_f are the deviations at the class marks of the classes just preceding the first frequency and following the last frequency, respectively, that is, the classes outside the actual distribution but adjacent to it. In applying this formula to an assumed origin the sum of the deviations must be corrected (Σd_c) as follows:

$$\Sigma d_c = \Sigma d + i(f_z - f_a) \div 24$$

After these formulas are solved, the computation of the standard deviation follows the usual short-cut method, that is:

$$\sigma_c^2 = \Sigma d_c^2/n - c^2$$

where $c = \Sigma d_c/n$. The process is illustrated in Example 20*a*.

Example 20*a*.—The corrected standard deviation, $\sigma_c^2 = \Sigma d_c^2/n - c^2$, where $\Sigma d_c^2 = \Sigma d^2 - i(ni + f_a d_p - f_z d_f)/12$; $nc = \Sigma d_c = \Sigma d + i(f_z - f_a)/24$; f_a and f_z are the first and last frequencies, respectively, and d_p and d_f are the deviations of the classes adjacent to the distribution, as indicated. The calculation may be carried through in units of the class interval (i).

	m	f	d	fd	fd^2
	1	0	$-4(d_p)$	0	0
	3	$2(f_a)$	-2	-4	8
$A_a =$	5	4	0	0	0
	7	3	2	6	12
	9	$1(f_z)$	4	4	16
	11	0	$6(d_f)$	0	0
	$i = 2$	$n = 10$		$\Sigma d = 6$	$\Sigma d^2 = 36$

$$\begin{aligned}
\Sigma d_c^2 &= \Sigma d^2 - i(ni + f_a d_p - f_z d_f)/12 \\
&= 36 - 2(20 - 8 - 6)/12 = 35. \\
\Sigma d_c &= \Sigma d + i(f_z - f_a)/24 \\
&= 6 + 2(1 - 2)/24 = 5.9167; \\
c &= \Sigma d_c/n = 5.9167/10 = 0.59167 \\
\sigma_c^2 &= \Sigma d_c^2/n - c^2 = 3.5 - 0.35007 = 3.14993. \\
\sigma_c &= \sqrt{3.14993} = 1.775
\end{aligned}$$

(as compared with $\sigma = 1.8$ by the usual method).

Sheppard's correction.—The method of correcting the standard deviation, as just described, made use of the formula,

$$\Sigma d_c^2 = \Sigma d^2 - i(ni + f_a d_p - f_c d_f)/12$$

in which the deviations (d) were taken from an assumed origin. If the deviations are considered to be taken from the arithmetic mean, this same formula, divided by n, expresses the second moment, as corrected for the slope of the frequencies, about the arithmetic mean; that is,

$$\sigma_c^2 = \Sigma d^2/n - i(ni + f_a d_p - f_z d_f)/12n$$

An examination of the derivation of this formula (cf. *Journal of the American Statistical Association*, December, 1929, pp. 349–366) will show that the terms $f_a d_p - f_z d_f$ are allowances made for the lack of what is called "close contact," that is, for the failure of the distribution to extend to an infinite number of frequencies of decreasing size at the extremes, as would be the case with a theoretical normal distribution. If these terms are dropped, the formula reduces to:

$$\sigma_c^2 = \Sigma d^2/n - i^2/12$$

where the deviations (d) are taken from the arithmetic mean. If the class interval (i) is taken as unity, the factor i^2 may, of course, be disregarded. This method of correcting the standard deviation is known as Sheppard's correction, and is theoretically correct for a frequency distribution which is entirely normal in form. It may therefore be applied to distributions which approximate this type. With ordinary distributions, however, the method as previously described is to be preferred if any correction for the slope of the frequency is to be made. Sheppard's correction, as applied to the data of Example 20*a*, gives

$$\sigma_c^2 = \Sigma d^2/n - i^2/12 = 3.24 - 0.333 = 2.907$$
$$\sigma_c = 1.705$$

where the deviations (d) are taken from the arithmetic mean. This method of correction gives apparently too small a result, as might be expected from the fact that it deducts for errors in non-existent classes at the extremes.

Corrected interpolation of percentiles.—If frequencies are assumed to be sloped toward the mode, it is evident that the interpolation of the median, quartiles, or other percentiles will be affected by this fact, especially if the slope is extreme. The result of such a correction is to shift the percentile to a point nearer the mode. As a result, the quartile deviation is somewhat decreased. Hence if rather precise results are required, it may sometimes be advisable to make corrections for the slope. This may readily be done, if the distribution is reasonably regular in form, by means of the formula for a corrected percentile (P_c):

$$P_c = m - f/b \pm [(f/b - i/2)^2 + 2i(n\% - \Sigma_1) \div b]^{1/2}$$

where the sign of the radical is determined by the sign of b; $n\%$ is n times the fraction represented by the percentile, as one-fourth for the first quartile, three-fourths for the third quartile, etc.; and b is defined as the frequency following the given class (f_{+1}) less the frequency preceding the given class (f_{-1}), divided by twice the class interval, $b = (f_{+1} - f_{-1}) \div 2i$. The other symbols have the same significance as in the interpolation of quartiles. If $b = 0$, this special formula is not required. As applied to the third quartile of Example 19 (p. 73), the interpolation, where $b = (1 - 4) \div 4 = - 0.75$, becomes:

$$\begin{aligned} Q_3 &= 7 + 4 - [(-4 - 1)^2 + 4(7.5 - 6) \div (-0.75)]^{\frac{1}{2}} \\ &= 11 - (25 - 8)^{\frac{1}{2}} = 6.88 \end{aligned}$$

Interpolating percentiles from the class mark (m).—The usual formula for interpolating percentiles may be changed algebraically into the form:

$$P = m + (i/2)\,(f_v - f_w) \div f$$

where

$$f_v = n\% - \Sigma_1 \quad \text{and} \quad f_w = f - f_v \text{ (or } \Sigma_2 - n\%);$$

that is, the percentile frequency (f) is intersected so as to divide the frequency column into the sub-totals required by the percentile, e.g., in the case of the median, into two equal parts.

By means of the above formula, a convenient method of finding the average deviation may be stated. The class marks are written as unit or step deviations from the median class (if the median is at a class limit either of the adjacent classes may be taken as the origin). The correction for the median class is

$$c_m = f_s(f_v - f_w) \div f$$

where f_s denotes the smaller of the two numbers, f_v and f_w. The deviations obtained by classes are totaled as positive. Dividing by n gives AD in step units, and this result times i is the average deviation as usually obtained from the median. The method is illustrated in the accompanying example. It is applicable only to the median as origin.

Example 21.—Average deviation by abbreviated step method,* median taken as

* If the corrections for class of origin and slope (c_1 and c_2) are to be made, they may be expressed for this special case as follows (for notation see Example 20, p. 80):

$$\begin{aligned} c_1/i &= 0.5\,[(f_v^2 + f_w^2)/f - |f_v - f_w|] \\ &= 0.5\,[(100 + 900)/40 - 20] = 2.5. \\ c_2/i &= [f - D(f_w - f_v)/2f] \div 6 - (f_a + f_z)/24 \\ &= [40 + 10(20)/80] \div 6 - (10 + 20)/24 \\ &= 42\tfrac{1}{2}/6 - 30/24 = 7.08\tfrac{1}{3} - 1.25 = 5.83\tfrac{1}{3}. \\ \Sigma' d'_c/i &= 75 + 2.5 - 5.83\tfrac{1}{3} = 71\tfrac{2}{3}. \\ \Sigma' d'_c &= 143\tfrac{1}{3}. \end{aligned}$$

origin; f_v is the part of f required to make the cumulatives equal $n/2$; and f_w is $f - f_v$; f_s indicates the smaller of the two parts, f_v and f_w; c_m is a correction applied to the fd/i column to allow for the deviations in the median class.

	m	f	d/i	fd/i
	3	10	-2	-20
	5	30	-1	-30
(MdC)	7	$40\begin{cases}10 = f_v\\30 = f_w\end{cases}$	0	$-5 = c_m$
	9	20	1	20
		$n = 100$		$\Sigma' d'/i = 75$

$c_m = f_s(f_v - f_w) \div f = 10(10 - 30) \div 40 = -5.$

$AD = (\Sigma' d') \div n = 2(75) \div 100 = 1.5.$

$Md = m + (i/2)(f_v - f_w) \div f = 7 + (1)(10 - 30) \div 40 = 6.5.$

The logarithmic standard deviation.—Distributions which are skewed in logarithmic form are sometimes measured by means of the geometric mean and the logarithmic standard deviation. The computation is made by substituting log m for each m, and finding the logarithmic average (log G) and the logarithmic standard deviation (log σ_r), as in Example 22. The antilogs of these measures are the geometric mean (G) and the standard deviation ratio (σ_r), respectively. The latter may be used as a measure of skewness. The reason that logarithmic normal distributions are thus measured is that when the m's are changed to a logarithmic scale they become normal distributions, hence the measures are more logically applied to the logarithmic scale. There is in this case a relation between the standard deviation ratio (σ_r) and the standard deviation as computed in the ordinary way, as follows,

$$\sigma = G[a(a - 1)]^{1/2}$$

where

$$a = \text{antilog } [(\log \sigma_r)^2/0.4343]$$

and the relation of the arithmetic mean to the geometric mean and the median and mode is as follows:

$$AM/a^{1/2} = G = Md = aMo$$

These equations hold exactly for logarithmic distributions which are entirely normal, but are approximately true of many skewed distributions which assume something like the logarithmic form.

If distributions approximating the logarithmic normal form are somewhat irregular, it may be better not to compute the geometric mean and logarithmic standard deviation directly because irregularities

in the data may produce erratic results. In effect, the distribution may be smoothed by finding the three quartiles and computing the required measures from these quartiles as follows:

$$\log \sigma_r = (\log Q_3 - \log Q_1) \times 0.7413$$

in which case G may be found by the formula:

$$\log G = (\log Q_1 + \log Q_3 + 1.2554 \times \log Q_2)/3.2554$$

Example 22.—The logarithmic standard deviation or standard deviation based upon log m instead of m. The geometric mean and the logarithmic standard deviation are logical measures of logarithmic normal distributions since such a distribution on a logarithmic base becomes normal about the geometric mean as mode.

m	f	$\log_m$	$f \log_m$	d (log)	d^2 (log)	fd^2 (log)
3	2	0.4771	0.9542	−0.2469	0.0610	0 1220
5	4	0 6990	2.7960	−0.0250	0 0006	0.0024
7	3	0.8451	2.5353	0.1211	0 0147	0.0441
9	1	0.9542	0 9542	0.2302	0.0530	0.0530
			10)7 2397			10)0.2215
			$\log G$ = 0.7240			$\log \sigma_r$ = 0.02215
			G = 5.30			σ_r = 1.0523

EXERCISES

1. (A) What are the methods of computing and the purposes of the measures of dispersion?

   (B) For each of the following distributions find the measures of dispersion: AD (from AM), σ (from AM), QD, and the coefficients of dispersion for each of these measures.

(a) m	f	(b) m	f	(c) m	f	(d) m	f	(e) m	f
2	3	4	2	1	1	2	1	2	1
3	5	6	4	2	3	4	2	4	4
4	6	8	6	3	5	6	5	6	6
5	4	10	5	4	2	8	3	8	4
6	2	12	3	5	1	10	1	10	1

(C) Find AD, σ, QD, and the coefficients of dispersion for the distributions given in Exercise 2 (c) and (d), p. 57.

2. Compare the average deviation with the standard deviation of the following items:

(a)	70;	20;	40;	10;	50;	30;	60.
(b)	140;	40;	80;	20;	100;	60;	120.
(c)	99;	101;	93;	105;	95;	107.	

3. Compare the average deviation from both the median and the mean with the standard deviation in the following distributions:

(a) m	f	(b) m	f
6	1	6	2
8	14	8	12
10	25	10	24
12	27	12	25
14	18	14	17
16	9	16	10
18	4	18	7
20	2	20	3

4. Compute the average deviation, standard deviation, and quartile deviation in the following distributions:

(a)		(b)		(c)		(d)		(e)		(f)		(g)		(h)		(i)	
m	f	m	f	m	f	m	f	m	f	m	f	m	f	m	f	m	f
5	1	6	2	3	20	2	10	8	30	4	4	10	3	4	4	6	2
15	4	18	4	5	50	4	40	10	60	8	7	20	7	8	6	10	5
25	3	30	3	7	40	6	50	12	50	12	5	30	6	12	5	14	6
35	2	42	1	9	10	8	20	14	20	16	3	40	3	16	3	18	4
										20	1	50	1	20	2	22	3

5. Calculate the arithmetic mean, median, average deviation, standard deviation, and quartile deviation of the following distributions:

(a) m	f	(b) m	f	(c) m	f	(d) m	f
2	3	3	1	3	1	2	2
4	9	5	6	5	7	4	5
6	10	7	7	7	11	6	7
8	8	9	7	9	9	8	6
10	6	11	4	11	6	10	5
12	3	13	3	13	4	12	3
14	1	15	2	15	2	14	2

6. Calculate measures of dispersion for the tabulations given under "Gathering and Presenting Data" and "Averages," using the corrections and special formulas in the more important cases. Check quartiles by cumulative charts.

7. The following table gives an abbreviated record of the percentage of net profits earned by certain representative American corporations (indicated by the letters A to K) during the years 1913 to 1924, inclusive.

Plot the lowest (Q_0), first quartile (Q_1), median (Q_2), third quartile (Q_3), and highest (Q_4) profit for each year (profit on Y-scale and time on X-scale). Connect the lowest profit in 1913 with the lowest in 1914, and so on to 1924. Similarly connect the first quartile points, the second quartile or median points, the third quartile points and the highest points, respectively, thus drawing five irregular but non-intersecting lines across the chart. These lines will indicate the general range and course of profits during the period covered (cf. Chart 13).

Corporation:

Year	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*
1913	4.5	2.6	0.5	3.9	8.0	6.8	12.1	9 3	16.2	11.0	21 2
1914	1.9	1.1	3.9	3.3	5.1	5.9	7.5	10.0	20.4	8.2	13.7
1915	7.9	0.6	3.1	5.0	6 7	11.1	4.4	21.1	12.7	26.7	15.3
1916	2.3	8.0	5 4	12 8	10.2	9.6	16.3	31.0	25.2	20.6	41.0
1917	8.4	6.1	2.3	12.1	14.3	11.2	15.9	25.3	32.8	20 6	17.8
1918	3 3	9.2	10.6	1.5	7.2	12.0	8.9	15.1	13.0	22.1	17.9
1919	9.4	4.6	0.5	6.2	10.7	8.3	16.0	14.1	19.1	23.4	12.5
1920	2.9	4.4	8.6	7.0	0.7	5.5	10.2	13.0	19.7	12.1	15.8
1921	−9.2	−6.1	−5.0	1.2	0.1	4.3	6.5	2.7	15.2	10.2	8.0
1922	1.8	6.8	−0.5	3.6	8.3	17.9	7.6	12.0	13.5	9.7	5.2
1923	6.9	1.5	3.0	4.6	8.3	15.0	6.0	7.5	12.0	11.7	10.0
1924	4.1	−0.5	2.6	6.2	9.5	12.2	5.5	6.7	8.0	14.5	11.0

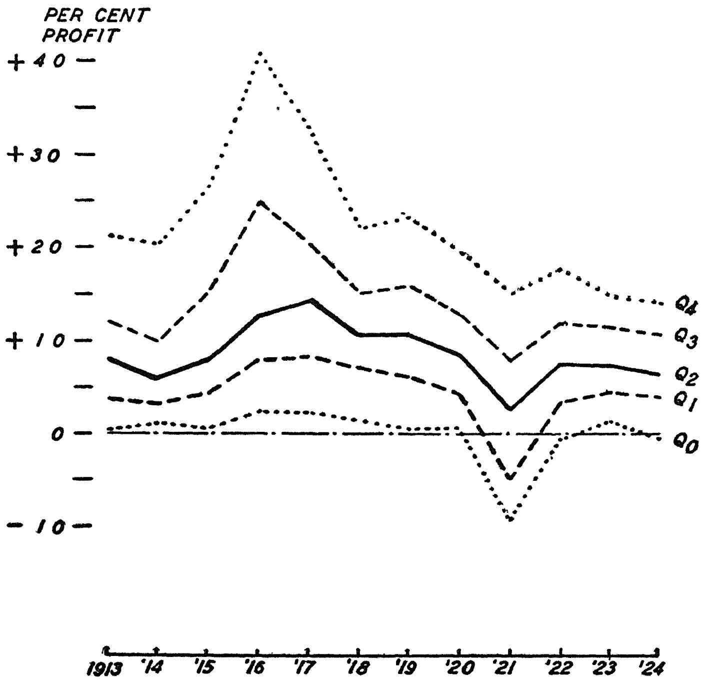


CHART 13

Quartiles and limits of profits earned by representative corporations, 1913–1924, as indicated by the data of Laboratory Exercise 7.

NOTES ON EXERCISE 7

The limiting items and quartiles are usually found by distributing the data in a frequency table, by recopying them as a simple array, or by merely ranking them in the original tabulation. The limiting items are the largest and smallest, which may be considered as the fourth and zero quartiles, respectively. The first quartile is the $(n + 1) \div 4$ or third item; the second quartile is the $(n + 1) \div 2$ or sixth item;

and the third quartile is the $3(n + 1) \div 4$ or ninth item. In this case, $n = 11$, the number of corporations. The five items in 1913 required for plotting are therefore: $Q_0 = 0.5$; $Q_1 = 3.9$; $Q_2 = 8.0$; $Q_3 = 12.1$; $Q_4 = 21.2$. If, however, the number of items were somewhat larger, a frequency tabulation for each year would be advisable. When the quartile positions are not integral, the fraction one-half is usually interpolated by taking the average of the two items lying directly above and below the position indicated. Other fractions are usually taken to the nearest unit. Thus, if $n = 16$, Q_1 is taken as the $(n + 1) \div 4 = 4.25$, or fourth item; Q_2 is taken as the $(n + 1) \div 2 = 8.5$, or average of eighth and ninth items; and Q_3 is taken as the $3(n + 1) \div 4 = 12.75$, or thirteenth item. Or, interpolation may be by spaces.

ANSWERS

1. (B)	*AM*	*Md*	*AD*	Coef.	σ	Coef.	*QD*	Coef.
(a)	3.85	3.833	0.98	0.255	1.195	0.310	0.925	0.242
(b)	8.30	8.333	1.96	0.236	2.390	0.288	1.85	0.222
(c)	2.92	2.900	0.78	0.267	1.037	0.356	0.667	0.235
(d)	6.17	6.200	1.56	0.252	2.075	0.336	1.33	0.211
(e)	6.00	6.000	1.50	0.250	2.000	0.433	1.500	0.214

	AM	*AD*	Coef. *AD*	σ	Coef. σ	*QD*	Coef. *QD*
(C) **Ex. (2c)**	5.2	2.048	0.3938	2.4	0.4615	1.8	0.3523
Ex. (2d)	4.4	1.024	0.2327	1.2	0.2727	0.9	0.2028

2.	*AD*	σ
(a)	17.14	20
(b)	34.29	40
(c)	4.33	5

3.	σ	$AD(R = Md)$	$AD(R = AM)$	*Md*	*AM*
(a)	2.912	2.292	2 24	11.7407	12
(b)	3.196	2.4896	2 5632	11.96	12.32
(a)	With c,	2.366	2.375		

4.	*AM*	*Md*	*AD*	*SD*
(a)	21.00	20.00	8.00	9.165
(b)	21.60	21.00	9.12	10.8
(c)	5.667	5.600	1.444	1.699
(d)	5.333	5.400	1.444	1.699
(e)	10.750	10.667	1.594	1.854
(f)	10.00	9.429	3.80	4.472
(g)	26.00	25.00	9.00	10.677
(h)	10.60	10.00	4.20	4.844
(i)	14.20	14.00	3.86	4.712

	Q_1	Q_2	Q_3	*QD*	Coef.
(a)	13.75	20.0	28.33	7.29	34.6%
(b)	13.5	21.0	30.00	8.25	37.9
(c)	4.4	5.6	7.0	1.3	22.8
(d)	4.0	5.4	6.6	1.3	24.5
(e)	9.33	10.67	12.20	1.433	13.3
(f)	6.57	9.43	13.20	3.32	33.5
(g)	17.86	25.00	33.33	7.74	30.2
(h)	6.67	10.00	14.00	3.67	35.5
(i)	10.40	14.00	18.00	3.80	13.4

5.

	AM	Md	AD	σ	QD
(a)	6.9	6.600	2.49	2.96	2.222
(b)	8.6	8.286	2.56	3.12	2.304
(c)	8.6	8.222	2.42	2.94	2.152
(d)	7.6	7.333	2.69	3.24	2.429

6. Dispersion for problems 5-a, 5-b, and 5-c in "Gathering and Presenting Data."

5–A—1923/1913

	I	II	III	IV	V
AM..........	147.64	140.25	199.37	189.33	155.68
σ............	40.81	32.73	34.73	36.42	33.25
σ_c............	40.5	32.4	34.3	36.4	33.1
Md...........	140.00	137.33	202.67	193.33	160.00
$AD(R = Md)$..	30.90	26.29	27.13	29.78	21.62
QD...........	28.18	22.74	23.35	25.42	14.82
Coef. QD......	0.19	0.16	0.12	0.13	0.09
$AD(R = AM)$.	......	......	......	......	23.66
Mo...........	135.7	123.0	212.0	200.0	162.0

	VI	VII	VIII	IX	Σ
AM..........	193.33	149.23	212.90	144.80	166.60
σ............	38.01	64.98	48.34	42.81	49.13
σ_c............	37.8	64.88	48.19	42.6	48.8
Md...........	190.00	125.45	215.00	142.86	163.44
$AD(R = Md)$..	31.11	48.34	38.54	29.36	39.33
QD...........	25.71	40.62	31.92	21.87	36.18
Coef. QD......	0.13	0.28	0.15	0.15	0.22
$AD(R = AM)$.	......	......	......	......	39.68
Mo...........	180.0	117.5	230.0	142.0	162.1

5–B—1924/1913

	I	II	III	IV	V
AM..........	155.36	143.21	192.92	182.67	148.65
σ *...........	54.15	31.12	34.02	47.23	31.71
Md...........	147.65	141.36	194.67	173.33	147.50
$AD(R = Md)$..	33.05	23.78	27.34	37.78	23.31
QD...........	18.64	20.80	24.00	35.00	16.71
Coef. QD......	0.12	0.15	0.12	0.19	0.11

	VI	VII	VIII	IX	Σ
AM..........	185.26	143.08	201.94	139.20	163.93
σ *...........	33.30	67.23	50.35	45.08	49.21
σ_c*...........	......	......	......	......	48.89
Md...........	185.00	121.67	192.50	137.14	159.10
$AD(R = Md)$..	26.05	48.93	40.89	31.77	36.73
QD...........	20.54	43.19	33.75	25.21	29.69
Coef. QD......	0.11	0.31	0.17	0.18	0.18
$AD(R = AM)$.	......	......	......	......	37.33
Mo...........	......	......	......	......	145.79

* Calculated on basis of $f\%$.

5–C—1925/1913

	I	II	III	IV	V
AM..........	165.33	157.53	191.04	182.67	147.57
σ............	43.12	34.79	34.10	56.57	30.25
Md...........	163.85	156.43	187.06	167.50	143.57
$AD(R = Md)$..	......	......	......	......	22.49
QD...........	......	......	......	......	16.61
Coef. QD......	......	......	......	......	.11
$AD(R = AM)$.	32.04	28.91	27.99	43.55	
Mo...........	171.8	......	......	......	......

	VI	VII	VIII	IX	Σ
AM..........	184.21	139.09	193.75	149.63	167.18
σ............	34.38	......	......	......	46.10
Md...........	182.00	118.57	184.29	145.00	165.57
$AD(R = Md)$..	......	......	......	......	35.84
QD...........	......	......	......	......	29.05
Coef. QD......	......	......	......	......	.18
$AD(R = AM)$.	......	......	......	......	35.97
Mo...........	......	......	......	......	160.75

6–A

	(1)	(2)	(3)	(4)	(5)
AM..........	100.42	98.07	103.59	96.28	95.34
σ............	14.42	12.55	7.98	8.07	13.81

7.

Year	Q_0	Q_1	Q_2	Q_3	Q_4
1913	0.5	3.9	8.0	12.1	21.2
1914	1.1	3.3	5.9	10.0	20.4
1915	0.6	4.4	7.9	15.3	26.7
1916	2.3	8.0	12.8	25.2	41.0
1917	2.3	8.4	14.3	20.6	32.8
1918	1.5	7.2	10.6	15.1	22.1
1919	0.5	6.2	10.7	16.0	23.4
1920	0.7	4.4	8.6	13.0	19.7
1921	−9.2	−5.0	2.7	8.0	15.2
1922	−0.5	3.6	7.6	12.0	17.9
1923	1.5	4.6	7.5	11.7	15.0
1924	−0.5	4.1	6.7	11.0	14.5

CHAPTER V

INDEX NUMBERS

ONE of the most useful devices employed by statisticians in the measurement of economic and social changes is a series of index numbers. The term index has a rather general meaning and may be applied to a single item or to a series of items. It is usually a percentage or a series of percentages expressing a comparison between the data for a certain month or year and another month or year or other period chosen as the base. For example, the Department of Labor Index of Wholesale Prices in the United States, having as its base the year 1926, gives index numbers for the years 1926 to 1931 as follows:

Year	Index number	Year	Index number
1926	100.0	1929	95.3
1927	95.4	1930	86 4
1928	96.7	1931	73.0

These figures mean that the general average of a large number of prices in the wholesale market, weighted in accordance with the estimated importance of each, declined during this five-year period so that a representative composite dollar's worth of goods costing one dollar in 1926 cost 73 cents in 1931. Such an index serves the purpose of giving a broad picture of what is occurring in wholesale markets, but like all averages, it may be misinterpreted because of irregularities in the various items included in the average. As a matter of fact, a few items rose during the same five-year period, some remained nearly constant, while many dropped more than the index indicates.

In a similar way, index numbers are computed for individual commodities, for limited groups of prices, for the volume of production in specific or general markets, and for various other purposes. In some cases, as in the measure of business activity, the indexes are expressed in terms of a computed normal which requires somewhat extended calculations, as will be discussed in a later chapter. The discussion in the present chapter is confined chiefly to index numbers in price and production fields, but the methods involved are applicable to social data as well.

Since somewhat complex formulas are required in connection with certain forms of index numbers, it may be well to state briefly the symbols and terminology commonly employed. Price percentages, such as those illustrated above, expressing changes in the prices of single commodities or groups of commodities, are called price indexes (P). Index numbers of the physical quantities produced or marketed, when based upon actual physical units, are called quantity indexes (Q). When physical production figures are expressed in terms of their cost at current prices, they are called value indexes (V). Value indexes, however, are not in themselves very significant, since they are influenced by changes in both the quantities and prices involved. Their chief use lies in connection with certain index number formulas of quantity and price.

Index numbers are sometimes expressed as mere aggregates, but as a rule they are expressed in percentages of the data in a base period more or less arbitrarily chosen. This percentage base is often the first year of the series, but it may be any year which is considered representative, or preferably it may be an average of several years. The choice of an adequate base is somewhat important in that if some items in the base year are at an abnormal level, comparisons may be distorted. The interval chosen as a percentage base is usually called merely the base, and is indicated by equating the time to 100, as in the expression $1926 = 100$. The term is, however, often applied to the period from which weights are chosen, but this period should be sharply distinguished from the percentage base, and is preferably designated as the weights base. Each year other than the percentage base year is designated as a " given " year. Index numbers may be computed for any convenient series of time intervals, as successive weeks, months, years, or decades. In the illustrations that follow, the year is used as the time interval, but the methods described may equally well be applied to months or other periods of time.

In formulas the specific quantities (number of bushels, yards, tons, etc.) are indicated by q; the prices at which they sell, p; the value of a specific commodity for the given year is therefore pq or v, and prices expressed as the amount-per-dollar are indicated by a. In practice, a and A (indexes of amounts-per-dollar) are not used except as a check on method. The aggregate value of a number of commodities marketed in a given period is Σpq. Subscripts may be attached to these symbols to indicate the time interval, as 0 for the base year, 1 or n for successive or any years, m for an average covering a group of years.

Simple index numbers.—Simple index numbers are merely percentages indicating successive changes in specific time series. Hence they

involve only ordinary percentage computations. They may be illustrated by the approximate data of production and average price of iron in the United States during the years 1919 to 1923, as given in Example 23.

Example 23.—Simple index numbers. Production and price of iron annually in the United States, 1919–1923, and annual value produced, together with index numbers derived as percentages, the year 1919 being taken as base. The original quantity, price, and value series are each divided by their respective base items in obtaining the percentage form. In each year the quantity index multiplied by the price index gives the value index. The reason for this is easily seen by reference to the data from which the index numbers are derived.

Year	Production and price of iron			Index numbers		
	Quantity (millions of tons)	Price (dollars)	Value (millions of dollars)	Quantity (%)	Price (%)	Value (%)
(Base) 1919..	30.6	32	979.2	100	100	100
1920..	36.4	44	1601.6	119	138	164
1921..	16.5	27	445 5	54	84	45
1922..	26.9	24	645.6	88	75	66
1923..	40.1	28	1122.8	131	88	115

Composite index numbers.—In the preceding section, simple index numbers covering a period of years were obtained expressing (*a*) relative changes in the physical quantities produced, that is, a quantity index; (*b*) relative changes in the prices at which the goods were marketed, that is, a price index; and (*c*) relative changes in the number of dollars-worths produced, that is, a value index. In practice it is the price index that is most commonly required, although indexes of quantity expressing the volume of business from time to time are often needed. As has been suggested, value indexes are not in common use, since they are too indefinite. For example, if we knew that the value output of a certain industry during a given year was 50% above that of the preceding year, we could not tell without further information whether to attribute the gain to an increase in physical volume, a rise in prices, or a combination of the two.

The chief difficulty encountered in computing composite index numbers of the market arises from incommensurability of the various units employed. For example, it may be necessary to condense into one composite figure the production of iron, the amount of transportation, the electric power production, etc., from year to year; or, again, a com-

posite may be required of the prices of goods entering into the laborer's standard of living. The difficulty that stands in the way of making such composites, whether of quantity or price, is that there is no simple way of adding such units as pounds, bushels, yards, car-loadings, and kilowatt-hours, or of weighting their prices. That is, the central difficulty in making composite index numbers lies in finding a satisfactory common denominator of the diverse physical units ordinarily employed.

Index numbers of value ($V_n = \Sigma p_n q_n / \Sigma p_0 q_0$).—The difficulty just mentioned does not, however, extend to the value index. If the relative changes in the number of dollars-worths of goods marketed is desired, all one has to do is to multiply the quantity and the price of each class of goods for each year specified, and add the various products. The total values in dollars thus found for each year may then be reduced to percentages relative to some year or other period taken as a base, just as in the case of simple index numbers. These percentages are the required index number of value (V). The method is illustrated in Example 24. The data are expressed in two forms. In the first form (I), the years are in columns and the commodities (A and B for tons of iron, bales of cotton or other products) are in rows. In the second form (II), the columns and rows are reversed.* For convenience, simple numbers have been chosen not representative of any particular line of production.

Example 24.—Index numbers of value ($V_n = \Sigma p_n q_n / \Sigma p_0 q_0$).

FORM I

Year	Commodity A			Commodity B			Σv	Value index (%)
	p	q	v	p	q	v		
(Base) 1900	1	8	8	2	6	12	20	100
1901	3	7	21	3	4	12	33	165
1902	2	9	18	4	8	32	50	250

FORM II

Commodity	1900			1901			1902		
	p	q	v	p	q	v	p	q	v
A	1	8	8	3	7	21	2	9	18
B	2	6	12	3	4	12	4	8	32
Σpq			20			33			50
$V_n = \Sigma p_n q_n / \Sigma p_0 q_0$			100 (%)			165 (%)			250 (%)

* Both forms are given throughout most of this chapter, so that the method will not be identified with either. In actual work, both forms, or modifications of them to suit the data at hand, are employed.

Aggregative index numbers of quantity and price.—The method most commonly used in calculating index numbers of quantity and price is called aggregative, because, like the method of computing a value index, it compares aggregate values for the given years. It is based on the following considerations. First, if it should happen that during certain years prices should remain constant while quantities varied, then the value index would obviously be in effect an index of quantity. Also if quantities should happen to remain constant while prices varied, then the value index would in effect be a price index. These considerations suggest that a quantity index may be calculated by the same procedure as a value index if prices are held constant at some representative figure; and likewise a price index may be calculated by similarly holding quantities constant. In either case the figures held constant are regarded as weights, stressing the relative importance of the particular variables and making possible the aggregating of the quantities and prices of dissimilar units. The typical prices or quantities used as weights are generally taken arbitrarily as of the percentage base period or other period regarded as typical or convenient for the purpose at hand. In practice they are often taken from the last census year, or from some other period when an adequate study was made. Sometimes this weights base period may be an average of the required data during two or more years. In Example 24*a* the census year 1900 is chosen as typical, and the constants are taken as of that year.* Since this is also the percentage base year the resulting indexes are said to be base-weighted. On this assumption, quantities and prices vary as indicated in Example 24*a*, but as will be shown later, the method is mathematically only an approximation. As in Example 24, the data are set down in two forms, in which the columns and rows are interchanged.

Example 24*a*.—Index numbers of quantity and price. Data of Example 24. Form 2 might be abbreviated, if several years were included, by omitting the column of constant p's and q's in all years except the first.

* Although the weights base is often identical with the percentage base, as in Example 24*a*, yet it is by no means necessary that this should be the case. In fact, the weights base might be a convenient census year or other period when complete data are available, prior to the period over which the index numbers are computed. From the theoretical point of view it is desirable that the weights base as well as the percentage base be designated in connection with a given series of index numbers.

FORM I

A. Index numbers of quantity, aggregative method, base-weighted

$$(Q_n = \Sigma p_0 q_n / \Sigma p_0 q_0)$$

Year	Commodity A			Commodity B			Σv	Quantity index (%)
	p	q	v	p	q	v		
(Base) 1900	1	8	8	2	6	12	20	100
1901	1	7	7	2	4	8	15	75
1902	1	9	9	2	8	16	25	125

B. Index numbers of price, aggregative method, base-weighted.

$$(P_n = \Sigma p_n q_0 / \Sigma p_0 q_0)$$

Year	Commodity A			Commodity B			Σv	Price index (%)
	p	q	v	p	q	v		
(Base) 1900	1	8	8	2	6	12	20	100
1901	3	8	24	3	6	18	42	210
1902	2	8	16	4	6	24	40	200

FORM II

A. Index numbers of quantity, as in Form I

Commodity	1900			1901			1902		
	p	q	v	p	q	v	p	q	v
A	1	8	8	1	7	7	1	9	9
B	2	6	12	2	4	8	2	8	16
Σpq			20			15			25
$Q_n = \Sigma p_0 q_n / \Sigma p_0 q_0$			100 (%)			75 (%)			125 (%)

B. Index numbers of price, as in Form I

Commodity	1900			1901			1902		
	p	q	v	p	q	v	p	q	v
A	1	8	8	3	8	24	2	8	16
B	2	6	12	3	6	18	4	6	24
Σpq			20			42			40
$P_n = \Sigma p_n q_0 / \Sigma p_0 q_0$			100 (%)			210 (%)			200 (%)

The aggregative method of computing a price index may be further illustrated by data from farm prices in Iowa, as issued by the State College at Ames. The index is expressed in percentages of pre-war prices, the base being taken as the average prices during 1910–1914. The typical quantity of each important commodity produced annually is estimated over several years in the earlier twenties. The actual quantities used as weights are not, however, a full year's production,

but are scaled down for convenience so that at base prices their aggregate value is \$100. That is, the actual quantities (q), as first tabulated, were multiplied by the prices in the base year (p_0q), and the total (Σp_0q) was divided into each of the original quantities (times 100 to change to percentages) to give the quantity weights. As a result of this adjustment, the sum of the column p_0q_w equals approximately 100. This is done for convenience in obtaining successive index numbers, since as a result the formula $P_n = \Sigma p_nq_w \div \Sigma p_0q_w$ reduces to $\Sigma p_nq_w \div 100\% = \Sigma p_nq_w$. The calculation for the year 1925 arranged in Form II, abbreviated, appears in Example 25.

Example 25.—Aggregative method illustrated. Index of prices of farm products in Iowa, 1925, on a 1910–1914 base using quantity weights (q_w) representing relative marketings, 1920–1924, in the units indicated. The quantity weights were obtained from the actual physical marketings (q) during the weights base period as follows: $q_w = 100q \div \Sigma p_0q$.

Aggregative method: $P_n = \Sigma p_nq_w \div \Sigma p_0q_w$.

Commodity	q_w	1910–1914		1925	
		p_0	p_0q_w	p_1	p_1q_w
Hogs	5.17 cwt.	\$7.30	\$37.741	\$11.08	\$57.284
Cattle	3.85 cwt.	6.39	24.602	8.43	32.456
Sheep	0.21 cwt.	4.51	0.947	7.48	1.571
Corn	24.98 bu.	0.53	13.239	0.86	21.483
Oats	19.12 bu.	0.35	6.692	0.39	7.457
Wheat	1.03 bu.	0.85	0.876	1.44	1.483
Hay	0.10 ton	9.82	0.982	11.23	1.123
Butter	40.62 lb.	0.25	10.155	0.41	16.654
Eggs	19.56 doz.	0.17	3.325	0.27	5.281
Poultry	14.58 lb.	0.10	1.458	0.18	2.624
		Σp_0q_w	\$100.017		\$147.416
	$P = \Sigma p_1q_w/\Sigma p_0q_w =$		100 (%)		147 (%)

Charting.—The charting of index numbers requires but little special comment. In general, index numbers are plotted as line charts with the time scale as the X-coordinate. Sometimes rectangles or bars are used to represent the ordinates.

An ordinary line graph plotted to the ratio scale is reproduced in Chart 14, which represents the course of wholesale prices in the United States for more than a century. The ratio scale is of advantage in such a chart in that it indicates the relative rate of change by the slope of the line; hence it is possible to estimate from the curve whether the

rate is increasing or decreasing during any given period. Chart 14*a* represents production over a period of years by a series of vertical bars, each of which is analyzed to show the relative uses of the product. Thus the chart may be read both as an index of business in the given industry, and, at the same time, as an analysis of the relation of that industry to other fields of business. Chart 14*b* is a percentage comparison of the United States with other countries in respect to certain important data; it illustrates a form of pictorial representation which is not

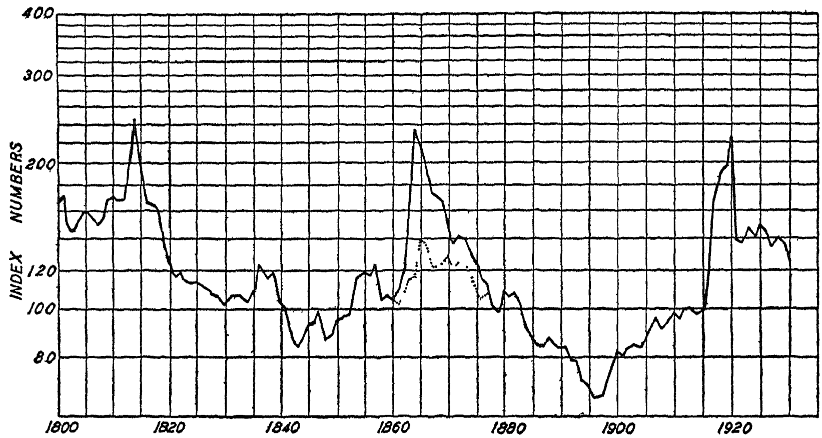


CHART 14

Annual index numbers of wholesale prices in the United States, 1800 to 1930. The dotted line from 1860 to 1878 represents prices reduced to a gold basis. The major inflation periods correspond to the War of 1812, the Civil War, and the World War. The longer swings of rising prices, as from 1843 to 1865 and from 1896 to 1914, were caused chiefly by gold production which was excessive relative to trade. Smaller irregularities generally represent business cycles. The figures are plotted to a ratio scale. The latter part of the series is the Department of Commerce index of wholesale prices reduced to a 1913 base. This is an aggregative index currently issued on a 1926 base. A weekly series on an extended basis has recently been begun.

subject to the error commonly involved in the use of surfaces or solids. Chart 14*c* is an example of a type of graphic representation commonly employed to represent stock market data. Small vertical bars indicate the range of prices during each day, and a line drawn through these bars forms a curve indicating the closing price day by day. At the bottom of the chart the volume of sales is represented by vertical bars. Such charts are extensively used by traders, who attempt to forecast the changes of the immediate future by reference to the previous cyclical movements, the volume of sales, and current happenings bearing upon the market.

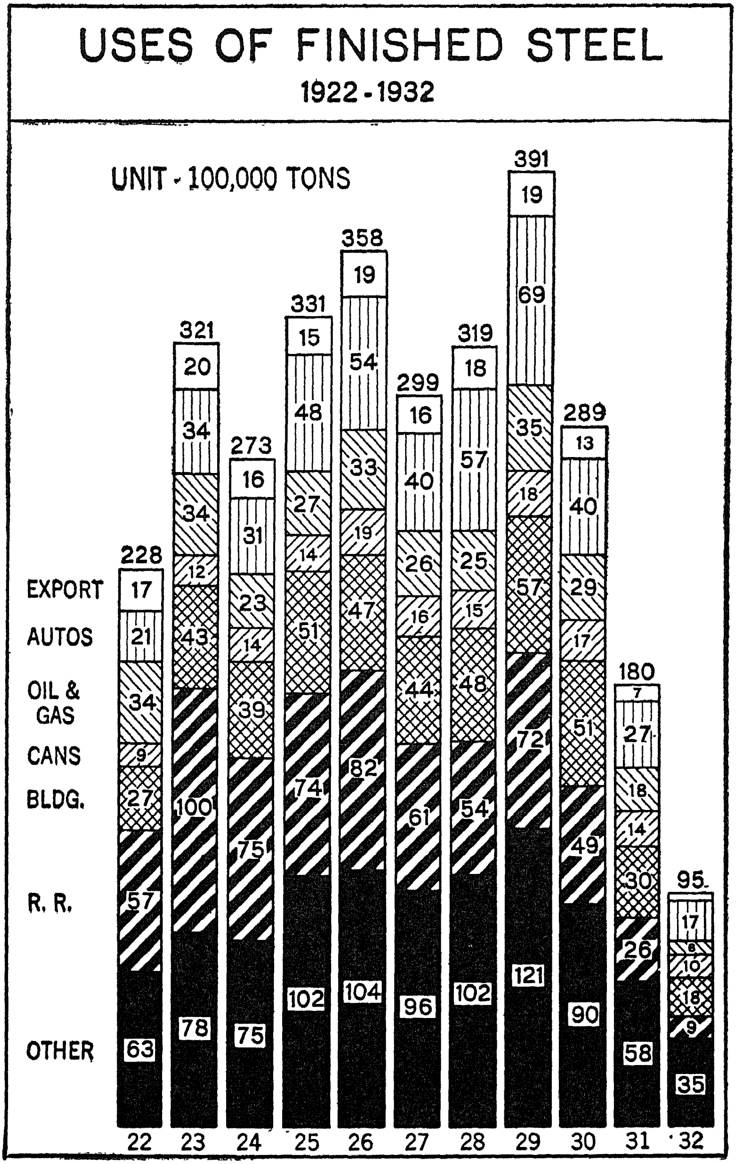


CHART 14*a*

Marketings of finished steel in the United States, 1922 to 1932 inclusive, classified by industries using the steel output. The chart is an illustration of one type of production index in which the output each year is classified according to use. Reprinted by permission from *Business Bulletin*, Cleveland Trust Company, January 15, 1933.

Corrected aggregative indexes.—Since the weights taken as constant in the common aggregative method are somewhat arbitrarily selected, the results obtained by this method will vary a little with the selection of the period from which the weights are taken, that is, with the choice of the weights base. In cases where data are complete, and where an exceptionally accurate result is desired, the aggregative method may be

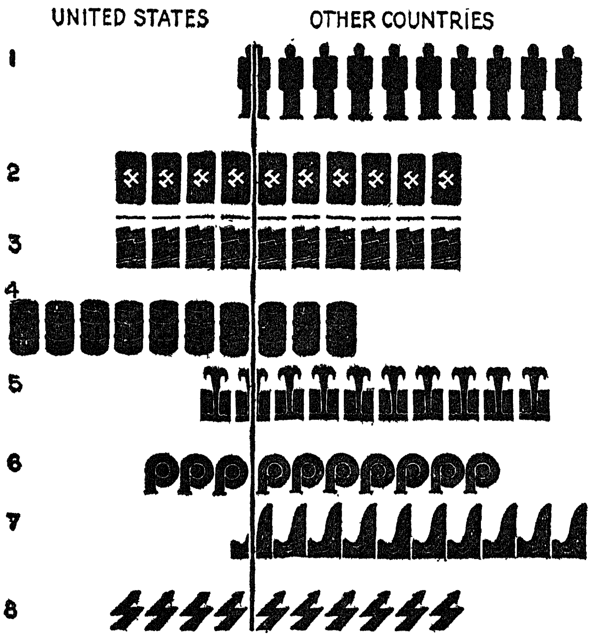


CHART 14*b*

Pictorial chart representing percentages of population, production, and resources in the United States and in other countries of the world. Each row of figures represents a total as 100%, and each individual figure represents 10% of this total. The successive rows represent (1) Population, (2) Coal production, (3) Coal resources, (4) Oil production, (5) Oil resources, (6) Developed water power, (7) Undeveloped water power, and (8) Electric power. Reprinted by permission from *Survey Graphic*, March 1, 1932, p. 627. Original by Dr. Otto Neurath, Gesellschafts und Wirtschaftsmuseum, Vienna.

corrected by a procedure which in effect takes into account the weights of each given year as well as those of the base year, as follows:

(a) Calculate the base-weighted index numbers of value (V), of quantity (Q_b), and price (P_b) according to the common aggregative method as explained in Examples 24 and 24*a*.

(b) Calculate a new quantity index (Q_r) by dividing the value index by the price index ($Q_r = V/P_b$); and calculate a new price index (P_r) by dividing the value index by the quantity index ($P_r = V/Q_b$).

(c) To obtain the final corrected quantity index (Q_c), average the two quantity indexes thus obtained ($Q_c = \overline{Q_b + Q_r}/2$); and similarly

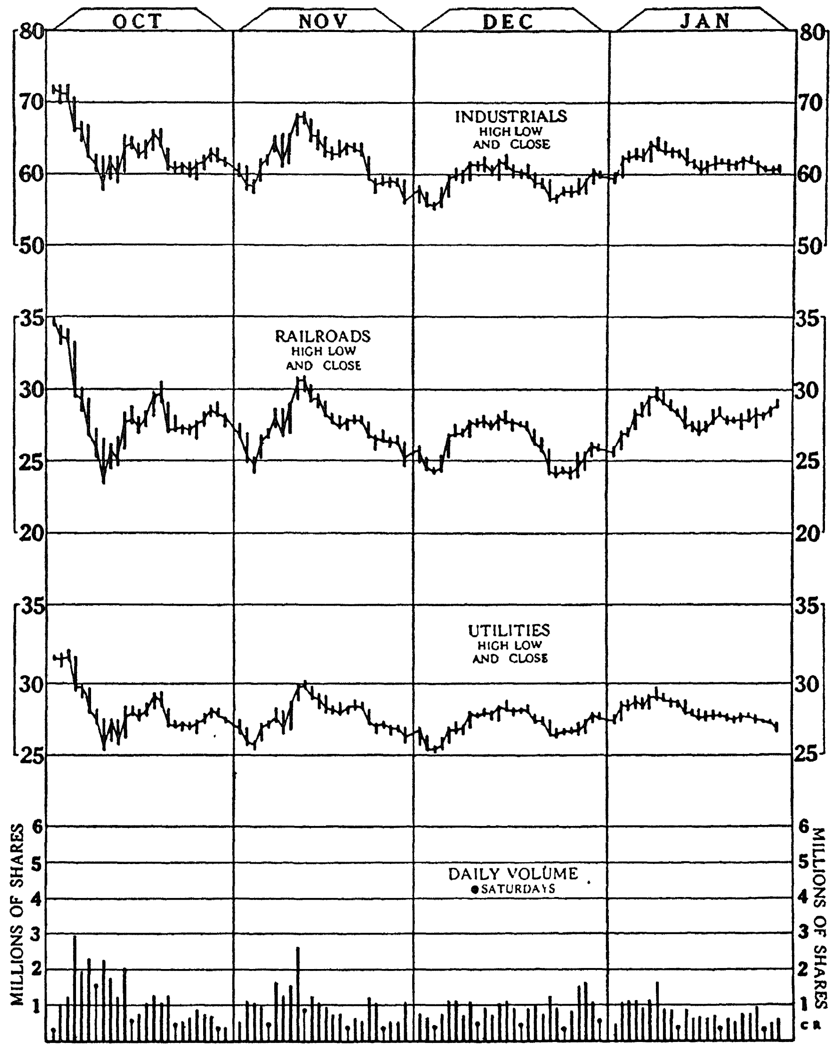


CHART 14c

Chart representing range of stock prices, indicating high, low, and close, together with volume of sales on the New York Stock Exchange, for October, November, December, 1932, and January, 1933. Stock prices are classified as industrials, railroads, and utilities. Reprinted by permission from *Wall Street Journal.*

to obtain the final corrected price index (P_c) average the two price indexes ($P_c = \overline{P_b + P_r}/2$), as illustrated in Example 26. The process

thus described is the same as that indicated by Fisher's ideal formula, to be described later, except that in the final averaging of Q_b and Q_r, and P_b and P_r, the arithmetic mean has, for convenience, been substituted for the geometric mean, which in this process it closely approximates.

The logic of the process just explained is in reality quite complex, but it may be somewhat superficially described as follows: Since axiomatically the price index times the quantity index should give the unequivocal value index, then on the assumption that the first quantity index (Q_b) is correct, the price index (P_r) obtained by dividing the quantity into the value index must also be correct. Hence Q_b and P_r are equally valid, as also for like reasons are P_b and Q_r. Sometimes it will happen that Q_r will equal Q_b and P_r will equal P_b, in which case the base-weighted results first obtained are confirmed. But as a rule, two different quantity indexes and two different price indexes will result. In each case the contrasting results may be taken to represent the outside limits between which the true results probably lie, on the assumption of the base year and the given year taken as alternate weights bases. The theoretical basis of the corrected aggregative method is discussed later in connection with Fisher's ideal method. It may be noted that almost the same results are obtained by averaging the weights instead of averaging the indexes.

Example 26.—The corrected aggregative method. The italicized figures have been calculated by the base-weighted aggregative method as given in Examples 24 and 24*a*; and $Q_r = V/P_b$, and $P_r = V/Q_b$. The final quantity index Q_c is the mean of Q_b and Q_r, and the final price index P_c is the mean of P_b and P_r. If the geometric mean were substituted for the arithmetic mean, the results would be identical with those obtained by Fisher's ideal formula.

Year	V	Base-weighted		Reverse-weighted		Average	
		Q_b	P_b	Q_r	P_r	Q_c	P_c
1900	*100*	*100*	*100*	100	100	100	100
1901	*165*	*75*	*210*	78.6	220	76.8	215
1902	*250*	*125*	*200*	125	200	125	200

The method of weighted relatives.—Before taking up certain inconsistencies inherent in the aggregative method, attention may be called to an alternative procedure known as the method of weighted relatives, or more briefly, the relative method. The method involves an attack upon the problem of composite index numbers at quite a new angle. It first computes simple index numbers, called in such a case "rela-

tives." Since these relatives are abstract, and are not affected by the type of unit from which they are derived, it might appear at first glance that the difficulty of combining dissimilar units had been solved. But in fact new difficulties arise. Obviously, averages of relatives must be weighted, but weights are not available that quite satisfy general mathematical considerations, as will be shown later. However, it would appear that the weights must represent the relative importance of the commodities in the market, and this importance is best represented by the value marketed during some period of time taken as typical. Certainly, mere quantities in tons and kilowatt-hours would not do as weights since they are incommensurable, and their influence would be altered by shifts in the physical unit, as from pounds to tons, or feet to yards. Hence typical values marketed are chosen as weights, and are applied to the relatives. The same weights are used for the price index as for the quantity index. If they are chosen from the base year the resulting index numbers are said to be base-weighted. In this case the results are the same as by the aggregative method, base-weighted, since for each commodity the relative times the value weight is identical with the product aggregated; that is $(q_1/q_0)p_0q_0 = p_0q_1$, and $(p_1/p_0)p_0q_0 = p_1q_0$. The method is illustrated for price indexes in Example 27. It will be noted that in Form I the weighted average of the relatives takes much the same form as in the aggregative method.

Example 27.—The method of weighted relatives, base-weighted by v of base year. The value index in Form I is repeated (cf. Example 24) to present the data and weights. Note that $\Sigma\%$ in the base year is the sum of the weights (Σw) expressed in percentages. Form II might be abbreviated, if several years were included, by omitting the columns of constant weights in all years except the first.

FORM I

1. Value index

	Commodity A			Commodity B				
Year	p	q	v	p	q	v	Σ	Value index
(Base) 1900	1	8	8 (wt)	2	6	12 (wt)	20	100
1901	3	7	21	3	4	12	33	165
1902	2	9	18	4	8	32	50	250

2. Quantity index

Weighted average of q-relatives; v_0 weights

	Commodity A			Commodity B				
Year	$q\%$	wt.	prod.%	$q\%$	wt.	prod.%	$\Sigma\%$	Quantity index, $\Sigma\%/\Sigma w$
(Base) 1900	100	8	800	100	12	1200	2000	100
1901	87.5	8	700	66.7	12	800	1500	75
1902	112.5	8	900	133.3	12	1600	2500	125

3. Price index

Weighted average of p-relatives; v_0 weights

	Year	Commodity A $p\%$	wt.	prod. %	Commodity B $p\%$	wt.	prod. %	$\Sigma\%$	Price index, $\Sigma\%/\Sigma w$
(Base)	1900	100	8	800	100	12	1200	2000	100
	1901	300	8	2400	150	12	1800	4200	210
	1902	200	8	1600	200	12	2400	4000	200

Form II

1. Value index—see above

2. Quantity index

Commodity	1900 $q\%$	wt.	prod. %	1901 $q\%$	wt.	prod. %	1902 $q\%$	wt.	prod. %
A	100	8	800	87.5	8	700	112.5	8	900
B	100	12	1200	66.7	12	800	133.3	12	1600
		20	)2000		20	)1500		20	)2500
			100			75			125

3. Price index

Commodity	1900 $p\%$	wt.	prod. %	1901 $p\%$	wt.	prod. %	1902 $p\%$	wt.	prod. %
A	100	8	800	300	8	2400	200	8	1600
B	100	12	1200	150	12	1800	200	12	2400
		20	)2000		20	)4200		20	)4000
			100			210			200

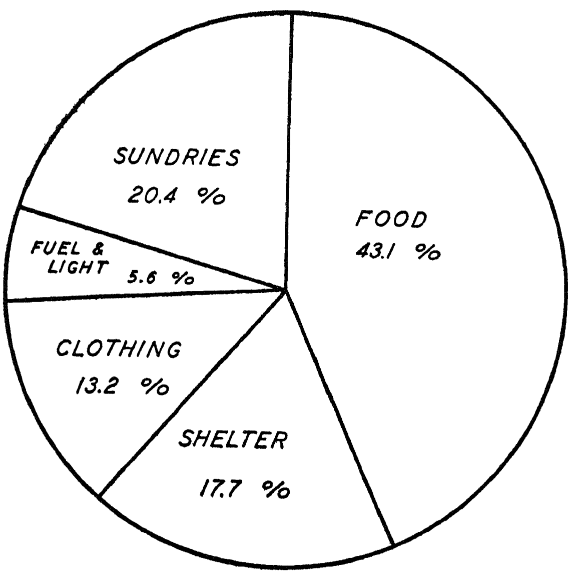


Chart 15

"Pie" chart of a typical cost of living budget, used as weights in calculating the cost of living (cf. Example 28).

The method of weighted relatives is usually more difficult to compute, and, since it has no decided advantages over the aggregative method, is not in common use. It often has a subsidiary use, however, as illustrated in Example 28, where group indexes obtained by the aggregative method are combined into a single index of the cost of living. A similar combination is made for each month for which the index is computed. The relative method is also illustrated for prices of farm products in Iowa (see Example 29) using the same data as in Example 25, with value weights representing the marketings in dollars, 1910–1914, reduced to percentages of the total marketings.

Example 28.—Cost of living, United States, June, 1927. Base year, 1913. Method of weighted relatives.

Commodities	Index	Per cent of budget	Product
Food	159	43.1	6,852.9
Shelter	169	17.7	2,991.3
Clothing	169	13.2	2,230.8
Fuel and light	160	5.6	896.0
Sundries	172	20.4	3,508.8
		100.0	)16,479.8
		Combined index:	164.8

Example 29.—Method of weighted relatives illustrated (cf. Example 25). Index of prices of farm products in Iowa, 1925, on a 1910–1914 base, using value weights (v_w) representing relative marketings in dollars worths during the weights base period, 1920–1924. Relative method: $P_n = \Sigma(p_1/p_0)v_w/\Sigma v_w$.

Commodity	v_w	p_0	p_1	$(p_1/p_0)\%$	$v_w(p_1/p_0)\%$
Hogs (cwt.)	35	7.30	11.08	151.8	5,313.0
Cattle (cwt.)	23	6.39	8.43	131.9	3,033.7
Sheep (cwt.)	1	4.51	7.48	165.9	165.9
Corn (bu.)	14	0.53	0.86	162.3	2,272.2
Oats (bu.)	6	0.35	0.39	111.4	668.4
Wheat (bu.)	1	0.85	1.44	169.4	169.4
Hay (ton)	1	9.82	11.23	114.4	114.4
Butter (lb.)	13	0.25	0.41	164.0	2,132.0
Eggs (doz.)	4	0.17	0.27	158.8	635.2
Poultry (lb.)	2	0.10	0.18	180.0	360.0
Σ	100				14,864.2
Index:	100				148.6

The geometric mean of relatives.—If the index of Example 29 is calculated backwards on a reversed base, that is, if average prices in 1910–1914 are compared with 1925 prices using the same relative method and the same value weights as before, an index of 68.134 (%) is obtained. This is larger than the reciprocal of 148.642 (%) (1 ÷ 1.48642 = 67.276%), and necessarily so, as was implied in an earlier chapter

(and cf. p. 313). Since it is just as legitimate to calculate the index "forward" as "backward," one or both of the indexes thus calculated are too large. The calculations using the arithmetic mean are therefore said to have an upward bias, though it is not possible without further investigation to say just how large this bias is. Obviously, the use of the geometric mean, which yields reciprocals when applied "forward" and "backward" will eliminate the bias, though this does not prove the geometric mean to be mathematically ideal. Hence the geometric mean is often used with the method of weighted relatives. The value weights and relatives are obtained as before; the only change lies in the use of the logarithms of the relatives, as previously explained in connection with the weighted geometric mean (cf. Examples 5 and 6, p. 39). The process as applied to the prices of farm products in Iowa in 1925 is illustrated in Example 30.

Example 30.—Weighted geometric mean of relatives illustrated (cf. Example 29). Index of prices of farm products in Iowa, 1925, on a 1910–1914 base, using value weights (v_w) representing relative marketings in dollars worths, 1920–1924. Geometric relative method: $\log P_n = \Sigma(\log \overline{p_1/p_0})v_w/\Sigma v_w$.

Commodity	v_w	p_0	p_1	$(p_1/p_0)\%$	$\log (p_1/p_0)\%$	$\log (p_1/p_0)\% \times$
Hogs (cwt.)....	35	7.30	11.08	151.8	2 1813	76.3455
Cattle (cwt.)...	23	6.39	8.43	131.9	2.1202	48.7646
Sheep (cwt.)....	1	4.51	7.48	165.9	2.2198	2.2198
Corn (bu.).....	14	0.53	0.86	162.3	2 2103	30.9442
Oats (bu.)......	6	0.35	0.39	111.4	2.0469	12.2814
Wheat (bu.)....	1	0.85	1.44	169.4	2.2289	2.2289
Hay (ton)......	1	9.82	11.23	114.4	2 0584	2 0584
Butter (lb.)....	13	0.25	0.41	164.0	2.2148	28.7924
Eggs (doz.).....	4	0.17	0 27	158.8	2.2009	8.8036
Poultry (lb.)....	2	0.10	0.18	180.0	2.2553	4.5106
Total........	100					100)216.9494
Index:	100					$\log P_n = 2.16949$
						$P_n = 147.7$

The chain index.—When it is not possible to obtain comparable data over a series of years, it is sometimes necessary to resort to a chain index. Such an index is computed from the first year to the second on the basis of what data are available, taking the first as the base. This may be done by any convenient method. The second and third years are then similarly compared. In the same way the comparisons may be carried through several years. The results are often called link-relatives; they are index numbers having the preceding year as the base. They may be chained by successive multiplication, to give ordinary index numbers. The index for the first year is 100, and

the index for each succeeding year is the product of the link-relatives down to and including that year.

A similar principle is employed in combining two index series that overlap for one or two years. The ratio of the average index number of one series to the corresponding average in the other for the overlapping years is found. One of the series, usually the later, is then multiplied by this ratio or its reciprocal so as to make the indexes average alike in the overlapping years. The results may be taken as continuous with the other series. In the overlapping years an average of the two results for each year may be taken as the final figure.

An index, whether obtained by chaining or combining as above, or by the more usual methods of calculation, may be changed to any required given base by merely dividing each index number by the index number for the period chosen as the base (cf. Examples 31 and 32). This obviously makes the index for the base period 100, and the other indexes remain proportional to it. It was formerly thought that this procedure was unjustified on the ground that the weights should be shifted when the base is shifted, but it has become evident that the choice of weights is not necessarily connected with the choice of a percentage base. If a series of index numbers is valid, it is just as valid when changed to another percentage base since it remains essentially the same series of ratios. Nevertheless the index numbers are conditioned by the weights base or bases, and for theoretical purposes these bases should be specified either directly or indirectly.

Example 31.—Chaining and combining indexes. It is assumed that link index numbers, as given in part A, have been computed for the years 1901 to 1905, each index being based on the preceding year. The chain is constructed from the links by successive multiplication, as indicated. If a different base is desired, as for example 1902, it may be obtained by dividing the series by the index for that year. The general principle involved is that, since index numbers in a series are ratios, they may be multiplied or divided throughout by any desired constant. The method is likely to be somewhat inaccurate because of inadequate data or incomplete comparability in the links or partial indexes.

A. Chaining the links, and shifting the base.

Year	Links (Base, preceding year)	Chain (Base, 1900)		Chain (Base, 1902)
1900	(100)	100	(÷107.1=)	93.4
1901	105 (×100.0=)	105	(÷107.1=)	98.0
1902	102 (×105.0=)	107.1	(÷107.1=)	100.0
1903	98 (×107.1=)	105.0	(÷107.1=)	98 0
1904	104 (×105.0=)	109.2	(÷107.1=)	102.0
1905	101 (×109.2=)	110.3	(÷107.1=)	103.0

B. Combining partial indexes (see also Example 32).

Year	Index 1	Index 2 (adjusted to link with 1)	Index 3 (adjusted to link with 2)	Combined index
1900	100			100
1901	110			110
1902	125	100 ($\times$125 = 125)		125
1903		95 ($\times$125 = 119)	92 ($\div$92 $\times$ 119 = 119)	119
1904			96 ($\div$92 $\times$ 119 = 124)	124
1905			100 ($\div$92 $\times$ 119 = 129)	129
1906			102 ($\div$92 $\times$ 119 = 132)	132

Example 32.—Splicing two index numbers. Department of Labor Index of Wholesale Prices, United States; A, computed on a 1913 base; and B, computed on a 1926 base. These indexes overlap in the year 1926, having at that date the ratio $A/B = 151/100 = 151$. Multiplying B by this ratio makes it continuous with A, or dividing A by this ratio would make it continuous with B. If the indexes overlap at two or more points, and have different ratios, the mean (preferably GM) of the ratios may be used to splice them and means of conflicting results taken. After splicing, the base may be changed as desired.

Year	A	B		B (spliced to A)
1913	100	...		
....	...	...		
....	...	...		
....	...	...		
1920	226	...		
1921	147	...		
1922	149	...		
1923	154	...		
1924	150	...		
1925	159	...		
1926	151	100	($\times$151/100=)	151
1927		95.4		144
1928		96.7		146
1929		95.3		144
1930		86.4		130
1931		73.0		110

Deflating.—The process of making allowance for the effect of changing price levels is called deflating. Thus, a record of bank clearings through successive years may be deflated by dividing each item by the corresponding index number of the general price level. Similarly, a wage series may be deflated by dividing by a cost of living index. In this case the results are called index numbers of real wages. In deflating it is necessary to make use of price indexes appropriate to the case in question. If the price index is representative of the values deflated, the result of the deflation is an index of quantity, that is, values at constant specific prices. If, however, the price index merely

reflects general price changes, the result of the deflation is the original values as they theoretically would have been if undisturbed by a changing price level; in other words, they are values at constant general prices. The deflation of wages, that is, the reduction of wages to real wages, is illustrated in Example 33. In this example the actual wage rate is divided by an index of the cost of living having as its base July, 1914. The resulting real wages are, therefore, expressed in dollars of July, 1914, purchasing power. They are theoretically proportional to the quantity of goods which could be purchased by the given wage, if no other variables were present. Such series or indexes, whether they have a specific base or not, are simply ratio series, and hence may be multiplied or divided by a constant and so reduced to any required base. When computed over a considerable period of time they may be misleading because of changing conditions such as altered standards of living which cannot be properly taken account of on a mathematical basis.

Example 33.—Deflating a value series for price changes. Wages per hour, common labor in road building, United States, divided by an index of the cost of living; base, July, 1914. The result is real wages in dollars of July, 1914, purchasing power. Real wages are then reduced to an index having the average real wage in 1923–1925 (average = $0.2317) as a base.

Year	Wages per hour	Index of cost of living; base, July, 1914	Real wages; dollars of July, 1914		Real wages; dollars of 1923–1925	
1920	$0.49	(÷) 197 (=)	$0.249 (÷0.2317=)		107.5	
1921	0 36	167	0.216		93.2	
1922	0.32	157	0.204		88.0	
1923	0 38	161	0.236	(Average = 0.2317)	101.9	(Base; average = 100)
1924	0.38	163	0.233		100.6	
1925	0.38	168	0.226		97.5	
1926	0.39	168	0.232		100.1	
1927	0.39	164	0.238		102.7	
1928	0.40	162	0.247		106.6	
1929	0.39	161	0.242		104.4	

Limitations on the use of index numbers.—Since index numbers of prices and quantities are averages, they are subject to the limitation of averages generally, and in addition are subject to certain special limitations arising from the field from which they are derived. Several recent writers, particularly European statisticians and economists, have emphasized these limitations so strongly that one would almost infer from their writings that the task of measuring economic change is futile.

(Cf. Haberler, "Der Sinn der Indexzahlen," Olivier, "Nombres indices de la variation des prix," and Keynes, "A Treatise on Money," Vol. I, Book 2.) Their criticisms have been extremely useful as a check against an exaggeration of the accuracy of the index numbers now obtainable. Perfect measurement is not attainable, even in the more exact sciences such as physics. Nevertheless, it is quite certain that the public will continue to view business and social change by means of indexes of price and quantity, inadequate though these may be, and that such indexes will be improved in the future.

Some of the limitations of index numbers arise out of difficulties in the theory itself. These difficulties will be considered in the supplementary discussion that follows in this chapter. But it may be said in advance that these difficulties are not really serious from a practical point of view. No one looks for absolute theoretical accuracy in index numbers; and as far as methods of calculation are concerned, sufficient accuracy for most practical purposes has already been attained. Greater degrees of accuracy may, however, be important in the future when ampler data are available, and when economic theory has worked out more fully its mathematical foundations. The question of theory is therefore one of prospective rather than immediate importance.

A second limitation on index numbers arises from the fact that complete data on which to base the calculations are so seldom available. Results are largely derived from sampling, and are therefore subject to a somewhat indefinite margin of error. Unfortunately the type of error which may be incurred is not the kind which can be calculated mathematically on the basis of probabilities. For example, the samples are commonly drawn from the price series most readily available, and these are likely to be the ones having fairly general and perhaps speculative markets. Hence they are not fully representative of many price series having less perfect markets. Only the more ample data of the future can give us a basis for fully evaluating the inadequacy of present results. However, in some fields, there are several indexes covering somewhat the same data, and, as a rule, these are in substantial agreement. It is therefore justifiable to rely tentatively upon such results.

Other limitations of index numbers arise from the wide range of variability of the data. For example, an index of wholesale prices extends over such a variety of commodities having different uses and subject to different influences that an average may not be significant. The remedy for this generality lies in a supplementary study of separate classes of commodities. This is being done in connection with the more important index series at the present time. The danger of extreme generality applies even more in relation to index numbers of the cost

would obviously have but little meaning. Hence cost of living indexes are usually confined to specific income subdivisions of the population; and the indexes in common use are those applicable chiefly to the working classes, since it is here that changes in the cost of living have the greatest social significance.

Many other limitations upon the use of index numbers might be mentioned, but they will come to the attention of the student in connection with later applications of statistics. This book is concerned chiefly with methods of analysis and principles of interpretation rather than with the many specific problems which arise in practical work.

SUPPLEMENTARY THEORY

A. Fisher's "Ideal" Formula

Test of consistency.—It has been seen that index numbers of quantity and price are conveniently constructed by computing values at constant typical prices and at constant typical quantities, respectively. Using the subscript w to indicate the prices or quantities thus held constant as weights, we have the formulas:

$$\begin{array}{lll} \text{Value index:} & V_1 = \Sigma p_1 q_1 / \Sigma p_0 q_0 \\ \text{Quantity index:} & Q_1 = \Sigma p_w q_1 / \Sigma p_w q_0 \\ \text{Price index:} & P_1 = \Sigma p_1 q_w / \Sigma p_0 q_w \end{array}$$

Indexes so constructed are mathematically valid only to the extent that the assumptions of typical weights are valid. That is, they answer the specific question relative to the changes in quantities or prices assuming the validity of the weights chosen. But in fact it is not easy to determine just what are typical weights. And it is not feasible to fall back on the simpler methods of averaging prices discussed in Chapter III, because incommensurable physical units must be combined. Hence certain axiomatic tests may be adopted in the light of which the formulas in common use may be criticized.*

* Fisher distinguishes three reversal tests which he denominates the preliminary test, the time reversal, and the factors reversal tests, respectively. The preliminary test concerns the interchange of any two commodities, and implies that they should be treated alike; the time reversal is the one here indicated as the reversal test; and the factor reversal is the one here indicated as the factors test. However, Fisher considers that the factor reversal test should not only meet the criterion $PQ = V$, but that also the p's and q's might be interchanged without invalidating the test. This implies that the same type of formula is to be used for quantities as for prices—an aspect of the factor reversal test which is here omitted on the grounds that, since price is a ratio of which quantities and values are the fundamentals, the formulas of price and quantities are dissimilar (cf. Chapter III). For the most part, however, Professor Fisher's analysis is followed. The so-called circular

The factors test.—As a test of the consistency of the q_w and p_w constants used as weights in the quantity and price formulas, we may set up the axiomatic criterion that, with specified data and base periods, the quantity index times the price index should equal the value index. This criterion may be called the factors test. Applying this test to the index numbers previously computed (cf. Example 26, p. 102), we have:

Year	P	Q	V
1900	100	100	100
1901	210	75	165
1902	200	125	250

But $P = 210$, times $Q = 75$, is $V = 157.5$ (percentages) rather than 165 as calculated, hence the price and quantity indexes are not consistent. In general it will be found that indexes thus computed with typical weights do not give consistent results though they may approximate the test and occasionally exactly meet it, as in the third year ($P = 200$, times $Q = 125$, is $V = 250$).

To study the foregoing test of consistency more closely it will be necessary to concentrate on data of two periods only, a base period and one given period. Let us assume that we are required to calculate a consistent set of indexes from such data. There is no question, of course, about the value index, since it does not require weighting. It is therefore necessary to inquire what typical weightings will give consistent indexes of quantity and price.

For illustration the data for the years 1900 and 1901 as previously used may be assumed. The value index is first computed by the usual method (cf. Example 34). The quantity and price indexes are also computed as before, taking the weights from the base year (a); but in addition they are also computed with weights taken from the second, or given, year (b). These two sets of weights may be taken to represent the extremes between which the weights typical of the two years should fall. It might be suggested that the typical weights should be taken as an average of the base and given weights, but experiment will show that no ordinary average will fulfill the criterion previously set up that $P \times Q$ should equal V.

Example 34.—Consistent index numbers of quantity and price. Data for commodities A and B, base year 1900 and given year 1901, including price (p), quantity (q), and value (v). Price and quantity indexes (P and Q) with base weights (a); and given weights (b); $Q_1 = \Sigma p_0q_1/\Sigma p_0q_0$ and $Q_1 = \Sigma p_1q_1/\Sigma p_1q_0$; also $P_1 = \Sigma p_1q_0/\Sigma p_0q_0$ and $P_1 = \Sigma p_1q_1/\Sigma p_0q_1$. When given weights are used, the results are said to be reverse-weighted.

I. Value index.

Year	Commodity A p	q	v	Commodity B p	q	v	Σv	Index (%)
1900	8	1	8	6	2	12	20	100
1901	7	3	21	4	3	12	33	165

II. Quantity index, base-weighted (a); and reverse-weighted (b).

Year	p	q	v	p	q	v	Σv	Index (%)
(a) 1900	8	1	8	6	2	12	20	100
1901	8	3	24	6	3	18	42	210
(b) 1900	7	1	7	4	2	8	15	100
1901	7	3	21	4	3	12	33	220

III. Price index, base-weighted (a); and reverse-weighted (b).

Year	p	q	v	p	q	v	Σv	Index (%)
(a) 1900	8	1	8	6	2	12	20	100
1901	7	1	7	4	2	8	15	75
(b) 1900	8	3	24	6	3	18	42	100
1901	7	3	21	4	3	12	33	$78\frac{4}{7}$

Consistent results.—It will be seen that Example 34 furnishes two sets of consistent results. Taking the subscripts b and r to indicate base-weighted and reverse-weighted, respectively, we may write:

$$P_bQ_r = V; \quad \text{or} \quad 75\% \times 220\% = 165\%$$

and

$$P_rQ_b = V; \quad \text{or} \quad 78\tfrac{4}{7}\% \times 210\% = 165\%$$

That results so taken are necessarily consistent may readily be seen by multiplying the formulas as given in Example 34. It will be seen that quantity, base-weighted, for the given year, is

$$Q_b = \Sigma p_0q_1/\Sigma p_0q_0$$

and price, reverse-weighted, for the same year, is

$$P_r = \Sigma p_1q_1/\Sigma p_0q_1$$

When these two expressions are multiplied, the term Σp_0q_1 cancels out, and the result is

$$P_rQ_b = \Sigma p_1q_1/\Sigma p_0q_0 = V$$

In the same way it may be shown that $P_bQ_r = V$.

Thus by using alternate base and given weights, two consistent sets of P and Q indexes are obtained, representing the limits between which the theoretically correct P and Q indexes for any given year relative to a base year may be assumed to lie. Hence the theoretically valid price index should be an average of P_b and P_r, and the theoretically valid quantity index should likewise be an average of Q_b and Q_r. The kind of average indicated is the geometric, since this average gives

results that meet the factors test. Using subscript i (ideal) to represent the means, we have

$$(P_b P_r)^{1/2} = P_i$$
$$(Q_b Q_r)^{1/2} = Q_i$$

and by regrouping the factors it is readily shown that $P_i Q_i = V$.

The aggregative index thus modified to make it consistent is known as Fisher's "ideal." Although Professor Irving Fisher did not originate this index, he did more than anyone else to investigate and evaluate the available methods of computation, and to show the desirability of the formula in question. His book, "The Making of Index Numbers," in which he sets forth the arguments for the "ideal" formula, is still the classic on the subject.

The base-reversal test.—Professor Fisher has suggested a second test of consistency applicable to quantity and price index numbers, known as the time reversal or base reversal test. It may be taken as axiomatic that if, in the problem of Example 34, the base should be changed to 1901, then the quantity index obtained for 1900 should be the reciprocal of the quantity index previously obtained; and the same reversal should hold for the price index. This test would obviously be met if the same weights were used in calculating forward (1900 base) and backward (1901 base). For example, the quantity indexes would be (the subscripts are used to represent the same years as before):

$$\text{Forward (1900 base)} \quad Q_1 = \Sigma p_w q_1 / \Sigma p_w q_0$$
$$\text{Backward (1901 base)} \quad Q_0 = \Sigma p_w q_0 / \Sigma p_w q_1.$$

The two indexes are obviously reciprocals, so that if $Q_1 = 200\%$, $Q_0 = 50\%$, as it evidently should. If, however, it were assumed that the weights ought to be chosen from the base year considered as the typical year, or year of reference, then the base reversal test would not necessarily hold for the ordinary aggregative method. It would, however, hold for Fisher's ideal method, as an inspection of the formula will show.

Applying the "ideal" index.—Fisher's "ideal" index is not very widely used, chiefly because it is usually difficult to gather all the required data. For example, suppose the Department of Commerce index of wholesale prices were to be constructed by the "ideal" method. It would then be necessary to gather not only monthly price figures for the several hundred commodities now included in the index, but also data on the monthly volumes of sales as well. This would greatly increase the task, and still further delay the calculations. Doubtless also the extra effort would yield better results if expended on additional price data. Besides, such an index at the best is only an approxima-

tion, since the data are never complete; hence an approximate formula may be justified by expediency. However, the ideal formula has a place with historical data and occasionally in current studies, and it may sometimes be useful as a check on the less accurate methods.

Computing the "ideal" index.—In actually computing the "ideal" index over an extended period of time, the procedure illustrated in Example 34 will prove very inconvenient. Since each year is compared directly with the base year, independently of other years, the reverse-weighted process calls for new weights for each year. It would therefore be necessary to rewrite the data of the base year with each successive given year, when using the given year weights. This tedious procedure may be avoided, however, by making use of the equations, previously given:

$$P_bQ_r = V, \quad \text{or} \quad Q_r = V/P_b$$
$$P_rQ_b = V, \quad \text{or} \quad P_r = V/Q_b$$

The base-weighted indexes may therefore be calculated by the usual procedure as illustrated in connection with the aggregative method (see Example 35), and the reverse-weighted indexes may then be obtained indirectly as indicated in the formulas just given; that is, the reverse-weighted quantity index is found by dividing the value index by the base-weighted price index for that year; and the reverse-weighted price index is similarly found. The procedure in thus obtaining the reverse-weighted indexes is identical with that followed in the corrected aggregative method (cf. Example 26, p. 102). The final quantity index for any year is theoretically the geometric mean of the two quantity indexes thus obtained for that year, though in practice the arithmetic mean is commonly taken as a close approximation. The final price index is similarly found by averaging the two price indexes.

Example 35.—Fisher's "ideal" method. The italicized figures are calculated by the base-weighted aggregative method as given in Examples 24 and 24*a* (pp. 94 and 95), and $Q_r = V/P_b$, and $P_r = V/Q_b$. Fisher's index (Q_i and P_i) is the geometric mean of the two preceding indexes of quantity and price, respectively. In practice, however, the arithmetic mean is often substituted for the geometric mean (cf. Example 26, p. 102).

		Base-weighted		Reverse-weighted		Average G.		Factors test
Year	V	Q_b	P_b	Q_r	P_r	Q_i	P_i	Q_iP_i
1900	*100*	*100*	*100*	100	100	100	100	100
1901	*165*	*75*	*210*	78.6	220	76.8	214.9	165
1902	*250*	*125*	*200*	125	200	125	200	250

B. General Index Number Theory

For theoretical purposes it is desirable to carry the analysis of index numbers as applied to general market changes one step further than is

done by Fisher's "ideal" formulas. To do this it will be necessary to inquire into the theory of that index with respect to the aggregating of incommensurable units.

Combining incommensurable units.—It is evident that any composite quantity index has in some manner combined incommensurable physical units (e.g., pounds, yards, kilowatt-hours, etc.). By some means not apparent on the surface, the incommensurable units have been rendered commensurable. The implicit theory of such a transformation in the case of the aggregative method is readily discovered. When a quantity is multiplied by the constant price taken as a weight (p_wq), it yields a result which may be interpreted as the number of revised physical units. The new physical unit is the quantity purchasable for one dollar at the standard price, as the constant price may be called. It is as if there were substituted for, let us say, a quart measure, another measure holding just the quantity costing one dollar at the standard price, and the goods were then remeasured by this new measure (a dollar's worth is $1/p_w$ quarts; and the number of such units in q quarts is $q \div \overline{1/p_w}$, or p_wq units). When this remeasuring of the physical units has been accomplished, they become commensurable, since each unit is a dollar's worth at standard prices. Hence the expression, Σp_wq, may be taken to mean the number of remeasured and commensurable physical units. It is true that some physical units could not actually be thus remeasured, but the theory is still applicable since constant multiples of the quantities may theoretically be substituted for the data.

Physical units in Fisher's index.—By beginning with the implied transformation of the physical units into commensurable dollars worths at standard prices, it is possible to express Fisher's "ideal" formula in accordance with the familiar principles of averaging as described in a previous chapter. The process is illustrated in Example 36. The physical quantities are reduced to units consisting of the amounts of the specified commodity purchasable for one dollar (a) in the base year (q/a_0 or qp_0), and the prices are adjusted to correspond (pa_0 or p/p_0).* Thus readjusted, the data relate to commensurable

* If prices in the given year are taken as standard, that is, if quantities are recomputed in terms of the amount-per-dollar in 1901, the following results are obtained:

Year	Average price	Average quantity	Average value	Average amount-per-dollar
1900	0.4615	16.25	7.5	2.1667
1901	1.0000	22.5	22.5	1.0000
Index 1901/1900	216.67	138.46	300.00	46.15

physical units (dollars worths) which may be treated as if they referred solely to pounds or solely to yards. The data for each year may now be averaged by weighting the ratios, as previously explained (p weighted with q and a weighted with v). The amounts-per-dollar are averaged as a check on the work. The indexes are obtained by taking the ratios of the averages in a given year (1901) to the averages in the base year (1900). The results meet the factors test ($PQ = V$), and the index of the amount-per-dollar is the reciprocal of the price index. Compared with the base-weighted and reverse-weighted aggregative indexes, the results are necessarily identical with the price reverse-weighted (P_r) and quantity base-weighted (Q_b) indexes.

Example 36.—Index numbers of price, quantity and value, adaptation of standardized quantities method, equivalent to common aggregative method, quantity (Q) base-weighted, and price (P) reverse-weighted.

I. Data of price (p), quantity (q), value ($pq = v$), and amount-per-dollar ($1/p = a$).

	p	q	v	a
		1900 (base)		
Commodity A	\$0.01	700 lb.	\$7	100 lb.
Commodity B	0.16	50 yd.	8	$6\frac{1}{4}$ yd.
		1901		
Commodity A	0.04	900 lb.	36	25 lb.
Commodity B	0.09	100 yd.	9	$11\frac{1}{9}$ yd.

II. Data remeasured in physical units taken as amount-per-dollar in base year; readjusted price = p times a of base year; readjusted quantity = q divided by a of base year; readjusted amount-per-dollar = a divided by a in base year. Averages q and v unweighted; averages p and a weighted. Indexes for 1901 compared with 1900 obtained as ratios of averages.

	p	q	v	a
		1900 (base)		
Commodity A	\$1.00	7 units	\$7	1 unit
Commodity B	1.00	8 units	8	1 unit
Average	1.00	$7\frac{1}{2}$	$7\frac{1}{2}$	1
		1901		
Commodity A	4.00	9 units	36	$\frac{1}{4}$ unit
Commodity B	$0.56\frac{1}{4}$	16 units	9	$1\frac{7}{9}$ units
Average	1.80	$12\frac{1}{2}$	$22\frac{1}{2}$	$\frac{5}{9}$
Indexes (%)........	180	$166\frac{2}{3}$	300	$55\frac{5}{9}$

Factors test: $PQ = V$; $180\% \times 166\frac{2}{3}\% = 300\%$.

The corresponding indexes obtained above by the use of base prices as standard were:

180.00	166.67	300.00	55.56

The geometric mean of each pair of index numbers is Fisher's "ideal" as follows:

197.48	151.91	300.00	50.64

Fisher's "ideal" index, therefore, is a compromise between the two results obtained by assuming, first, that the base prices are standard, and second, that the given prices are standard. This is not the same as assuming that the average prices of the two years are standard.

If Example 36 is recomputed using prices in the given year as standard (readjusted physical units = a_1) the results will be found to be identical with the aggregative base-weighted price index and the reverse-weighted quantity index. Thus we reach the important conclusion that, in terms of the aggregative method:

(a) If base year prices are taken as standard
P_r and Q_b are correct.
(b) If given year prices are taken as standard
P_b and Q_r are correct.

It is customary among some European writers to designate a price index number for a given year obtained by the base-weighted common aggregative method by the symbol beta (β), and the same type of index, reverse-weighted, by the symbol gamma (Γ).* Using these symbols to apply to both price and quantity indexes, we may state the foregoing conclusion as follows: Beta quantities and gamma prices imply standardizing the quantities in terms of the data of the base year, and gamma quantities and beta prices imply standardizing the quantities in terms of the data of the given year. Each pair of results is correct according to the point of view adopted. Fisher's "ideal" index may therefore be described as a compromise between the results obtained on the basis of these two diverse points of view. Since, as a general thing, there is logically no way of determining which year is standard, this compromise is plausible. It has one drawback, however; it does not readily lend itself to the calculation of consistent index numbers for three or more years; that is, it fails to meet the so-called circular test. This test sets up the hypothesis that index numbers derived from closely related data should be consistent for all dual comparisons; for example, if the index of the second year is twice that of the first year and the index for the third year is twice the second year, then the index for the third year should be four times that of the first year. Fisher's index will fail to meet this test since in each dual comparison the weights are shifted; that is, the point of view or "frame of reference" is changed.

A standard price.—An obvious alternative to the ideal index is an index computed on the assumption of a standard price derived from a longer time interval. In the case of a dual comparison this may be done by using as the standard price for any commodity the average

* The symbols are generally applied to price indexes but may appropriately be extended to quantity indexes. The beta price index is known as Laspeyres' formula and the gamma price index as Paasche's formula.

price of the two periods. For three or more periods, similarly, the average price of each commodity might also be chosen as standard unless for specific reasons some more restricted interval could be regarded as more typical.

But even though we might assume that the standard price should be the average price over the whole period covered by an index series (standardized quantities method), the question of the kind of average to use would still remain. The weighted arithmetic mean would at first thought appear to be the most logical since this mean balances the deviations and makes comparisons of aggregates correspond to comparisons of averages; also it is the mean which is used in compiling the data within each period of time. When this average is used the total values and the total quantities over the whole interval covered by the index series are equated. But, on the other hand, it might be argued that the standard price should be such that if a general change in the whole price level in one period is assumed (for example, if all prices in one year are arbitrarily doubled), the effect of this change should appear in the price index only, and not in the quantity index. If this requirement is set up, then the average employed in obtaining the standard price must be an unweighted geometric mean of the prices of each commodity in each year. However, this requirement does not appear to be absolutely essential since obviously in any case the results obtained are conditioned upon the assumption of a standard price. The problem becomes one of relativity, and the fact that the change in prices carries over into quantity need not, therefore, be considered in itself entirely abnormal.

A further consideration of a very general nature points in the direction of a standard price obtained by arithmetic means. As Walsh has argued, if all ratios of money to goods and goods to money in every unitary trade in a system of exchanges are included, the average should be unity. That is, the price level in any market considered as a universe is always unity, except as the ratios of all goods to one standard good are isolated, as in the ordinary measurement of prices. This consideration indicates the weighted arithmetic mean as preferable.

The relativity of index numbers.—But in any case, as King has so well shown from another point of view, the measurement of market change by index numbers is always relative to the given problem at hand. As the physicists have proved, the relativity of magnitudes applies even to physical measurements. Index numbers, therefore, are at best to be taken as conditioned by the specific problem involved

and by the "frame of reference" set up in the choice of a standard price. Whatever method may be used in determining the standard price (p_m), the formulas for the standardized quantity method, which assumes that some sort of an average standard price has been used, are as follows: *

$$P_n = (\Sigma p_n q_n / \Sigma p_m q_n) \div (\Sigma p_0 q_0 / \Sigma p_m q_0)$$

$$Q_n = \Sigma p_m q_n / \Sigma p_m q_0$$

$$V_n = \Sigma p_n q_n / \Sigma p_0 q_0$$

$$A_n = (\Sigma p_m q_n / \Sigma p_n q_n) \div (\Sigma p_m q_0 / \Sigma p_0 q_0)$$

The application of these formulas to simple data, where the standard price of each commodity is taken as the unweighted geometric mean, is illustrated in Example 37.

Example 37.—Index numbers of value (V), quantity (Q), and price (P) for two commodities (A and B) for the two years 1900 (base) and 1901, by the standardized quantities method. The standard price (p_m) of each commodity is the unweighted geometric mean (G) for the two years. Form I is somewhat abbreviated for convenience of computation; Form II is arranged to bring out more clearly the theory involved. In the second part of Form II, the data are written in readjusted form expressing the number of standard physical units and the prices of these units. The average of the price relatives is not expressed but would give the same result as that here indicated (cf. "Averaging Double Ratios," p. 51). The physical units of commodities A and B may be assumed to be pounds and yards, respectively.

* The standardized quantities method may be expressed entirely in terms of the averaging of relatives, and results identical with those here given will be obtained provided that price relatives are appropriately treated as double ratios. The problem takes the following form (cf. Example 37):

	Double relatives p_1/p_0	Weights q_1/q_0	Products v_1/v_0	wt.$_q$	prod.	wt.$_v$	prod.
Commodity A...	400	128.57	514.29	14	18	7	36
Commodity B...	56.25	200	112.5	6	12	8	9
				20	)30	15	)45
	P = 200				Q = 150		V = 300

Price relatives are weighted with quantity relatives, and the products are value relatives. Since the weights and products are themselves relatives, they must be given secondary weights in combining them. The secondary weights are the standardized quantities and the values in the base year. The average price relative is the ratio of the averages of the fundamental relatives.

FORM I

A. Value index. $V_n = \Sigma p_n q_n / \Sigma p_0 q_0$.

	Year	Commodity A p	q	v	Commodity B p	q	v	Σv	Index (%)
(Base)	1900	\$0.01	700 lb.	\$7	\$0.16	50 yd.	\$8	15	100
	1901	0.04	900	36	0.09	100	9	45	300
		G=\$0.02			G=\$0.12				

B. Quantity index. $Q_n = \Sigma p_m q_n / \Sigma p_m q_0$.

	Year	p	q	v	p	q	v	Σv	Index (%)
(Base)	1900	0.02	700	14	0.12	50	6	20	100
	1901	0.02	900	18	0.12	100	12	30	150

C. Price index. $P_n = (\Sigma p_n q_n / \Sigma p_m q_n) \div (\Sigma p_0 q_0 / \Sigma p_m q_0)$.

	(a)	V	÷ Q	= P	P (%)	or (b) V (%)	÷ Q (%)	= P (%)
(Base)	1900	15	20	0.75	100	100	100	100
	1901	45	30	1.50	200	300	150	200

FORM II

A. Data in original units.

	1900 p	q	v	a	1901 p	q	v	a
Commodity A	\$0.01	700 lb.	\$7	100 lb.	\$0.04	900 lb.	\$36	251 lb.
Commodity B	0.16	50 yd.	8	6 yd.	0.09	100 yd.	9	$11\frac{1}{9}$ yd.

B. Data expressed in standardized units.

	1900 p	q	v	a	1901 p	q	v	a
Commodity A	0.50	14	7	2	2.00	18	36	$\frac{1}{2}$
Commodity B	$1.33\frac{1}{3}$	6	8	$\frac{3}{4}$	0.75	12	9	$1\frac{1}{3}$
Average...	0.75	10	$7\frac{1}{2}$	$1\frac{1}{3}$	1.50	15	$22\frac{1}{2}$	$\frac{2}{3}$
Index, 1901/1900					200(%)	150(%)	300(%)	50(%)
Index, 1900/1901..	50(%)	66.7(%)	33.3(%)	200(%)				

Need for further research.—Some recent writers on index numbers have expressed the opinion that it is impracticable to attempt to develop a logical index number formula. This opinion is doubtless justified for most of the current work in index numbers, inasmuch as such work is based upon incomplete and sometimes inaccurate data. But there is another point of view to be considered. There is now developing a restatement of economics in mathematical and statistical form, which promises to be an important contribution to social science. In this process of development one of the central problems will inevitably be the trending of aggregated quantities and average prices. Hence, although a theoretically accurate index number formula may have no immediate practical value, it has a theoretical and prospective value that should not be overlooked.

The solution here presented in the form of the standardized quantity

method is not assumed to be final.* We are still in the stage of experimentation and theorizing. We have gone far enough to see, however, that the large majority of the diverse weightings presented by Professor Fisher in his classical work may be discarded, not only on

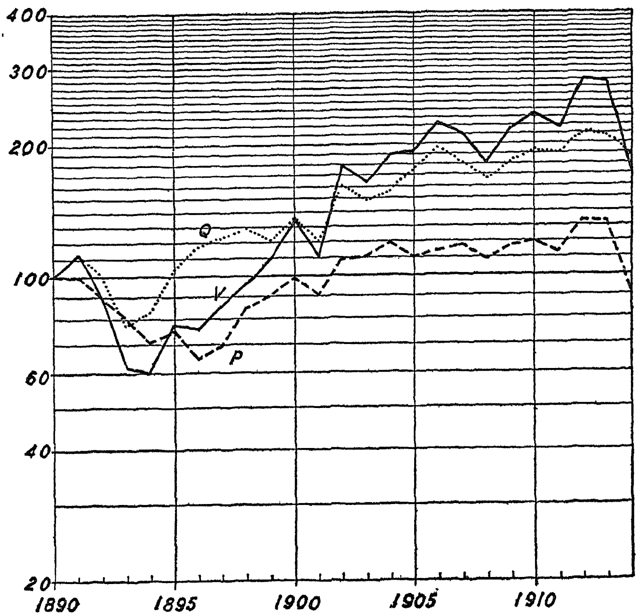


CHART 16

Index numbers of value (V), quantity (Q), and price (P) by Fisher's "ideal" formula using the data of Exercise 1 (cf. Laboratory Exercises at the close of this chapter). These indexes are plotted to the ratio scale so that as measured from the 100% line on any ordinate $Q + P = V$; that is, the factors test is met.

the practical basis which he there set up, but also on theoretical grounds. For example, an harmonic mean of ratios weighted by the type of unit represented by the denominator is scarcely defensible. We have also come to see that we must make a clear-cut distinction between the

* An objection that has been raised to the standardized quantity method is that, when the circular test is met, occasionally a situation will arise where all prices decline and yet the price level rises, or *vice versa*. The fact, however, that such a situation may arise is not at all a condemnation of the method. As was seen in connection with the averaging of double ratios, a parallel may easily be found in the averaging of double ratios arising out of physical data. For example, the speed of each runner might increase from one day to the next and yet the general level of speed might decline because of weightings in favor of the poorer runners, or *vice versa*. The parallel in price index numbers, when buying swings from one class of goods to another, is obvious.

averaging of ratios and the averaging of fundamentals. But much investigation must still be carried on before we can speak with certainty.

EXERCISES

1. Using the following data, compute index numbers of value, and base-weighted, reverse-weighted, and "Ideal" index numbers of quantity and price. Make a ratio chart of the results by the "ideal" method.

Data: Annual production and prices of three important commodities (*A*, *B*, and *C*), United States, 1890–1912. (Quantities in millions; prices in dollars; rough approximations only.)

	Year	*A*		*B*		*C*		Year	*A*		*B*		*C*	
		q	*p*	*q*	*p*	*q*	*p*		*q*	*p*	*q*	*p*	*q*	*p*
(Base)	1890	5 @	4	10 @	2	10 @	1	1903	7	4	15	2	16	1.5
	1891	6	4	11	1	11	2	1904	8	5	16	1.5	15	2
	1892	5	2.8	10	1.6	10	1.5	1905	9	4	17	2	18	1.5
	1893	4	2	7	2	9	1	1906	10	3.4	18	2	21	2
	1894	4	3	8	1.5	10	0.6	1907	8	5	20	1.4	19	2
	1895	5	2.6	11	2	10	0.4	1908	9	3	16	3	16	1
	1896	6	3	12	1.5	10	0.2	1909	9	4	20	1.9	17	2
	1897	6	2	13	1	12	1 5	1910	10	3	20	2.1	18	2.5
	1898	7	3	14	1	10	1.3	1911	9	2	19	2	20	2.6
	1899	6	3	11	2	15	1	1912	10	5.2	21	2	23	2
	1900	7	4	14	1.5	12	1.5	1913	10	5.2	21	2	22	2
	1901	6	5	12	1	13	1	1914	9	2	18	1.5	20	2
	1902	8	5	17	2	15	1							

2. (a) Using the following data, compute index numbers of value, quantity, and price by the common aggregative method, base-weighted.

(b) Compute the reverse-weighted indexes of quantity and price, and check by the factors relation.

(c) Compute Fisher's "ideal" index numbers.

(d) Compute index numbers by the standardized quantities method, but use the arithmetic mean instead of the geometric in finding typical prices.

Assumed data of production and price of three commodities, United States, 1915–1918. (Quantities in thousands of tons; prices in dollars.)

		Commodity *A*			Commodity *B*			Commodity *C*				Index
	Year	*q*	*p*	*v*	*q*	*p*	*v*	*q*	*p*	*v*	Σv	*V*
(Base)	1915	2 @	1	= 2	5 @	2	= 10	4 @	2	= 8	20	100
	1916	2	2		6	3		5	4			
	1917	1	2		6	5		6	3			
	1918	3	3		7	6		5	3			

3. Using the following data, compute index numbers of value, quantity, and price by the standardized quantities method (dissimilar physical units; prices in dollars), taking p_m as a geometric mean.

(a)

	Commodity A			Commodity B		
Year	q	p	v	q	p	v
1900	600	0.04	24	80	2.25	180
1901	3600	0.25	900	30	4.00	120

(b)

	Commodity A			Commodity B		
Year	q	p	v	q	p	v
1900	500	0.04	20	250	0.16	40
1901	1000	0 09	90	300	0.25	75

(c)

	Commodity A			Commodity B		
Year	q	p	v	q	p	v
1900	100	0.01	1	150	0.16	24
1901	3000	0.04	120	20	0.25	5

4. Using the following data, compute index numbers of value, quantity, and price by the method of weighted relatives, using as weights the values in the base year. Also find Fisher's "ideal" index numbers. Chart on ratio paper as a check (quantities in tons; prices in dollars).

Data, with designations as above, but for the years 1900–1904.

(Base)	1900	5 @ 4 = 20		12 @ 2 = 24		3 @ 2 = 6		50	100
	1901	6	3	15	3	6	2		
	1902	4	5	15	6	6	4		
	1903	6	6	18	7	6	5		
	1904	5	5	12	6	3	1		

5. Combine the following link-relatives into a chain index having 1900 as the base year:

Years, 1901–1904; link-relatives, 80, 130, 125, 90.

6. Combine the two index series given below into a single series having 1890 as the base year:

	Year	Index A	Index B
(Base A)	1890	100	
	....	...	
	1911	150	
	1912	147	98
(Base B)	1913		100
	1914		110
	1915		108

7. (a) Find the purchasing power of one dollar in wholesale markets, relative

to the year 1913, for the years as indicated (that is, deflate a dollar, using the accompanying index).

Year	Unit of value	Index of wholesale price level
(Base) 1913	$1	100
1914	1	98
1915	1	101
1916	1	127
1917	1	177
1918	1	194
1919	1	206
1920	1	226
1921	1	147
1922	1	149
1923	1	154

(b) Similarly compute the relative purchasing power of a dollar by years and by months since 1926, on the basis of Fisher's, the *Annalist*, and the Department of Commerce indexes of wholesale prices. Also compute the relative changes in the purchasing power of a dollar in retail markets as measured by an index of cost of living (cf. *Survey of Current Business*, last annual number).

8. (a) Using the following data, deflate bank debits for years indicated. Chart the results together with the debits as given.

Year	Bank debits outside New York (millions of dollars)	Index of general price level (per cent of 1913)
1913	75,181	100
1914	72,227	100
1915	77,253	103
1916	102,275	117
1917	129,540	139
1918	153,817	157
1919	211,175	173
1920	241,596	193
1921	191,941	163
1922	199,509	158
1923	225,330	165

(b) Obtain and deflate similar data of bank debits, retail sales, exports, etc., from *Survey of Current Business* and other sources.

9. Using the following data, compute an index of real wages for the years indicated. Plot data and results on ratio paper.

Year	Index of wages	Index of cost of living	Year	Index of wages	Index of cost of living
1913	100	100	1919	193	183
1914	102	102	1920	232	208
1915	104	104	1921	207	182
1916	118	111	1922	201	168
1917	134	131	1923	220	171
1918	168	159			

10. From the data of Exercises 2 and 4, compute index numbers of quantity and price using the relative method with the geometric mean of the relatives.

11. From the following four sets of assumed data of quantity (q) and price (p), compute index numbers of value (V), quantity (Q_b) and price (P_b) base-weighted, quantity (Q_r) and price (P_r) reverse-weighted, quantity (Q_i) and price (P_i) Fisher's ideal (arithmetic means), and the weighted geometric means of the quantity relatives (Q_g):

(a)

	Commodity A		Commodity B		Commodity C	
Year	q	p	q	p	q	p
1910	2	3	3	5	2	2
1911	4	3	6	3	3	3
1912	3	4	6	4	5	1
1913	2	6	5	6	4	2

(b)

	Commodity A		Commodity B		Commodity C	
Year	q	p	q	p	q	p
1910	4	1	5	2	3	2
1911	6	3	6	3	6	1
1912	8	5	7	2	6	3
1913	2	4	5	6	3	4

(c)

	Commodity A		Commodity B		Commodity C	
Year	q	p	q	p	q	p
1900	1	2	2	2	1	4
1905	2	$2\frac{1}{2}$	3	2	$1\frac{1}{2}$	4
1910	$1\frac{1}{2}$	4	2	3	2	3
1915	2	2	4	2	2	4

(d)

	Commodity A		Commodity B		Commodity C	
Year	q	p	q	p	q	p
1900	2	2	4	2	8	1
1905	3	3	6	2	6	$1\frac{1}{2}$
1910	2	$1\frac{1}{2}$	6	3	4	2
1915	4	1	8	1	4	2

12. Using the following data, construct percentage quantity weights on a 1913 base ($q_w/\Sigma p_0 q_w$), and find an aggregative index number of prices, base year, 1913. (Results will vary 1 to 10 points from the index of 22 items.)

Retail food, United States, consumption (q_w), in 1918 taken as typical. Prices, 1913–1923, in mills

Item	q_w	1913	1914	1915	1916	1917	1918	1919	1920	1921	1922	1923
Steak	32 lb.	223	236	230	245	290	369	389	395	344	323	335
Ham	22 lb.	269	273	261	294	382	479	534	555	488	488	455
Butter	66 lb.	383	362	358	394	487	577	678	701	517	479	554
Milk	337 qt.	089	089	088	091	112	139	155	167	146	131	138
Bread	531 lb.	056	063	070	073	092	098	100	115	099	087	087
Potatoes	704 lb.	017	018	015	027	043	032	038	063	031	028	029
Index, 22 items		100	102	101	114	146	168	186	203	153	142	146

13. Using the following data of farm prices in Iowa, and value and quantity weights, construct index numbers of prices by the common aggregative, the relative, and geometric relative methods.

	1910–14	1915–19	1920–24	1925	1926	1927	1928	1929	1930
Hogs....	7.304	12.562	8.680	11.08	11.62	9.49	8.61	9.48	8.80
Cattle...	6.388	9.668	7.650	8.43	7.98	8.92	10.85	10.78	9.24
Sheep...	4.510	8.022	6.052	7.48	6.84	6.57	6.94	6.49	4.64
Corn....	0.528	1.074	0.718	0.86	0.60	0.74	0.81	0.77	0.69
Oats....	0 346	0.540	0.402	0.39	0.34	0.41	0.43	0.39	0.33
Wheat..	0.852	1.668	1.244	1.44	1.28	1.22	1.09	1.07	0.79
Hay....	9.822	13.648	12.658	11.23	13.98	13.69	12.05	10.66	9.34
Butter..	0.254	0.376	0.410	0.41	0.42	0.44	0.46	0.46	0.36
Eggs....	0 168	0.278	0.262	0.27	0.28	0.23	0.25	0.26	0.19
Poultry..	0 098	0.158	0.176	0.18	0.20	0.18	0.198	0.191	0.154

1931

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Hogs....	7.10	6.40	6.80	6.80	6.10	5 40	6.00	5.90	5 20	4.40	4.10	3.40
Cattle....	7.90	7.10	6.90	7.00	6.30	6.00	6.10	6.30	6.30	6.10	6.60	5.40
Sheep...	3.80	3.80	4.00	4.10	3.60	3.40	2.50	2.50	2.20	2.30	2.50	2.20
Corn....	0.56	0.51	0.50	0 48	0.47	0.45	0.47	0.43	0.37	0.28	0.35	0.32
Oats....	0.27	0.26	0.26	0 26	0.24	0.23	0.21	0.17	0.17	0.17	0.21	0.21
Wheat..	0.63	0.62	0.63	0.63	0.63	0.60	0.40	0 37	0.37	0.37	0.48	0.43
Hay....	9.60	8.60	8.10	8.20	8.20	8.00	8.20	8.20	8.20	7.80	7.90	8.60
Butter..	0.29	0.27	0.29	0.28	0.24	0.23	0.24	0.25	0.28	0.33	0.31	0.30
Eggs....	0.18	0.11	0.17	0.151	0.115	0.119	0.125	0.149	0.150	0.186	0.232	0.220
Poultry..	0.148	0.138	0.149	0.151	0.133	0.145	0.145	0 159	0.152	0.123	0.132	0.125

1932

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Hogs....	3.40	3.20	3.80	3.40	2.70	2.50	4.20	3.90	3.55	3.00	2.80	2.40
Cattle...	5.00	4.40	4.80	4.80	4.30	4.30	5.90	5.90	6.00	5.30	5.00	4.10
Sheep...	2.00	2 20	2.60	2.60	2 00	1.90	1.90	1.90	2.00	2.00	1.95	1.65
Corn....	0.32	0 29	0.29	0.27	0.25	0.23	0.25	0.25	0.21	0.14	0.13	0.12
Oats....	0.21	0 21	0.21	0.21	0.19	0.17	0.14	0 12	0.12	0.098	0.10	0.10
Wheat..	0.42	0.42	0.43	0.43	0.41	0.39	0.34	0.37	0.37	0.35	0.34	0.32
Hay....	8.20	8.70	9.20	10.00	9.80	8.60	7.10	7.00	6.40	5.90	5.80	5.70
Butter..	0.27	0.22	0.22	0.21	0.19	0.17	0.17	0.20	0.20	0.19	0.20	0.22
Eggs....	0.141	0.107	0.083	0.088	0.096	0.086	0.099	0.130	0.143	0.20	0.238	0.269
Poultry..	0.121	0.107	0.111	0.109	0 105	0.087	0.10	0.108	0.11	0.091	0.086	0.075

Weights of specified commodities based upon monthly and annual income (value weights), and annual quantity weights; weights base, 1920–1924

Commodity	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug	Sept	Oct.	Nov	Dec.	Yr.		Q wt.
Hogs......	37	41	37	33	32	31	35	34	29	32	39	39	35		5.166
Cattle......	25	23	23	27	25	21	21	20	20	21	22	22	23		3.852
Sheep......	1	1	1	1	1	1	1	1	1	1	1	1	1		0.212
Corn	17	14	16	9	9	15	13	13	18	16	11	14	14		24.980
Oats.......	5	5	5	5	5	5	5	8	10	8	5	4	6		19.121
Wheat.....	1	1	1	1	1	1	1	3	2	2	2	1	1		1.030
Hay.......	1	1	1	1	1	1	1	1	1	1	1	1	1		0.101
Butter.....	9	10	11	13	16	18	17	15	15	14	13	11	13		40.621
Eggs.......	1	2	4	9	9	6	5	4	3	3	1	1	4		19.559
Poultry....	3	2	1	1	1	1	1	1	1	2	5	6	2		14.575

14. On the basis of the following quantities and prices, compute the indexes V Q_b P_b Q_r P_r Q_i P_i. Also compute an index of prices, using the relative method weighted with base year values—the weight being reduced to percentages totaling 100.

	Commodities					
	A		*B*		*C*	
Year	*q*	*p*	*q*	*p*	*q*	*p*
(Base) 1900	10	10	10	5	5	10
1901	12	7	8	7	7	8
1902	15	8	12	6	8	10
1903	11	8	12	4	6	10
1904	12	9	14	3	7	12
1905	13	10	10	4	6	14
1906	14	12	8	5	8	12
1907	14	10	8	6	8	14
1908	15	12	12	7	10	12

15. Plot the individual indexes and superimpose the combined index of business activity as worked out by the *Annalist* for the years 1923 to 1931. (See pp. 6 and 7, *Survey of Current Business*, 1932 Annual Supplement, United States Department of Commerce.) For a brief statement of the method and weights used in determining the combined index, see p. 288, *Survey of Current Business*, as above.

16. Deflate the wages of common labor in road building in the United States by the cost of living index and compute the monthly relatives using the monthly average for the year 1923 as 100%. For data on wages in road building, see p. 70, *Survey of Current Business*, 1932 Annual Supplement, and p. 22 of the same publication for cost of living index.

17. Tabulate local data of cost of living during recent years and calculate index numbers.

ANSWERS

1. Index numbers of value (V), quantity (Q), and price (P) by aggregative base-weighted and reverse-weighted methods, and their averages (Fisher's method). Figures in parentheses are values of specific commodities.

Year	V $\Sigma p_1q_1/\Sigma p_0q_0$	Base-weighted		Reverse-weighted		Average	
		Q $\Sigma p_0q_1/\Sigma p_0q_0$	P $\Sigma p_1q_0/\Sigma p_0q_0$	Q V/P	P V/Q	Q (arithmetic mean)	P
(B) 1890	(20–20–10) 100	(20–20–10) 100	(20–20–10) 100	100.0	100.0	100.0	100.0
1891	(24–11–22) 114	(24–22–11) 114	(20–10–20) 100	114.0	100.0	114.0	100.0
1892	(14–16–15) 90	(20–20–10) 100	(14–16–15) 90	100.0	90 0	100.0	90.0
1893	(8–14–9) 62	(16–14–9) 78	(10–20–10) 80	77.5	79.5	77.7	79.8
1894	(12–12–6) 60	(16–16–10) 84	(15–15–6) 72	83.3	71 4	83.7	71.7
1895	(13–22–4) 78	(20–22–10) 104	(13–20–4) 74	105.4	75.0	104.7	74.5
1896	(18–18–2) 76	(24–24–10) 116	(15–15–2) 64	118.7	65.5	117.4	64.8
1897	(12–13–18) 86	(24–26–12) 124	(10–10–15) 70	122.8	69.3	123.4	69.7
1898	(21–14–13) 96	(28–28–10) 132	(15–10–13) 76	126.3	72.7	129.2	74.4
1899	(18–22–15) 110	(24–22–15) 122	(15–20–10) 90	122.2	90.2	122.1	90.1
1900	(28–21–18) 134	(28–28–12) 136	(20–15–15) 100	134.0	98.5	135.0	99.3
1901	(30–12–13) 110	(24–24–13) 122	(25–10–10) 90	122.2	90.2	122.1	90.1
1902	(40–34–15) 178	(32–34–15) 162	(25–20–10) 110	161.8	109.9	161.9	110.0
1903	(28–30–24) 164	(28–30–16) 148	(20–20–15) 110	149.1	110 8	148.5	110.4
1904	(40–24–30) 188	(32–32–15) 158	(25–15–20) 120	156.6	118.9	157.3	119.5
1905	(36–34–27) 194	(36–34–18) 176	(20–20–15) 110	176.3	110.2	176.2	110.1
1906	(34–36–42) 224	(40–36–21) 194	(17–20–20) 114	196.5	115.5	195.3	114.8
1907	(40–28–38) 212	(32–40–19) 182	(25–14–20) 118	179.6	116.4	180.8	117.2
1908	(27–48–16) 182	(36–32–16) 168	(15–30–10) 110	165.4	108.3	166.7	109.2
1909	(36–38–34) 216	(36–40–17) 186	(20–19–20) 118	183.0	116.1	184.5	117.0
1910	(30–42–45) 234	(40–40–18) 196	(15–21–25) 122	191.8	119.4	193.9	120.7
1911	(18–38–52) 216	(36–38–20) 188	(10–20–26) 112	192.8	114.9	190.4	113.5
1912	(52–42–46) 280	(40–42–23) 210	(26–20–20) 132	212.1	133.3	211.0	132.7
1913	(52–42–44) 276	(40–42–22) 208	(26–20–20) 132	209.1	132.7	208.6	132.4
1914	(18–27–40) 170	(36–36–20) 184	(10–15–20) 90	188.9	92.4	186.5	91.2

2. (a)

Year	V	Q_b	P_b
1915	100	100	100
1916	210	120	175
1917	250	125	205
1918	330	135	240

(b)

Q_r	P_r
100	100
120	175
121.95	200
137.5	244.49

(c)

Year	Q_i	P_i
1915	100	100
1916	120	175
1917	123.47	202.48
1918	136.24	242.21

(d)

Q	P
100	100
119.4	175
122.2	204.2
136.1	237.5

3. (a)

Year	V	Q	P
1900	100	100	100
1901	500	150	333

(b)

V	Q	P
100	100	100
275	150	$183\frac{1}{3}$

(c)

Year	V	Q	P
1900	100	100	100
1901	500	200	250

4. (a)

Year	V	Q_b	Q_i	P_b	P_i
1900	100	100	100	100	100
1901	150	132	131.8	114	113.8
1902	268	116	119.5	218	224.5
1903	384	144	146.4	258	262.3
1904	200	100	100	200	200

5. 100, 80, 104, 130, 117

6. 100, ..., 150; 147; 150; 165; 162

7. 100; 102; 99; 79; 56; 52; 49; 44; 68; 67; 65

8. 75,181; 72,227; 75,003; 87,415; 93,194; 97,973; 122,066; 125,179; 117,755; 126,272; 136,564

9. 100; 100; 100; 106; 102; 106; 105; 112; 114; 120; 129

10.

Year	Q	P
1915	100	100
1916	119	173 2
1917	120	199.3
1918	134	227.4

Year	Q	P
1900	100	100
1901	130.0	108.3
1902	110.6	201.3
1903	142.1	239.5
1904	100	170.5

11. (a)

Year	V	Q_b	P_b	Q_r	P_r	Q_i	P_i	Q_g
1910	100	100	100	100.0	100.0	100.0	100.0	100.0
1911	156	192	84	185.7	81.3	188.8	82.6	191.0
1912	164	196	88	186.4	83.7	191.2	85.8	191 3
1913	200	156	136	147.1	128.2	151.6	132.1	151.8

(b)

Year	V	Q_b	P_b	Q_r	P_r	Q_i	P_i	Q_g
1910	100	100	100	100 0	100.0	100.0	100.0	100.0
1911	210	150	150	140.0	140 0	145 0	145.0	146.3
1912	360	170	195	184.6	211.8	177.3	203.4	167.3
1913	250	90	290	86.2	277.8	88.1	283.9	87.1

(c)

Year	V	Q_b	P_b	Q_r	P_r	Q_i	P_i	Q_g
1900	100	100	100	100.0	100.0	100.0	100 0	100.0
1905	170	160	105	161.9	106.2	161.0	105.6	158.9
1910	180	150	130	138.5	120 0	144.2	125.0	143.1
1915	200	200	100	200.0	100 0	200.0	100.0	200.0

(d)

Year	V	Q_b	P_b	Q_r	P_r	Q_i	P_i	Q_g
1900	100	100	100	100.0	100.0	100.0	100.0	100.0
1905	150	120	130	115.4	125.0	117.7	127.5	113.7
1910	145	100	155	93.5	145.0	96.8	150.0	89.1
1915	100	140	110	90.9	71.4	115.4	90.7	114.9

12. Indexes: 100; 103.2; 103.7; 117.0; 151.5; 165.3; 182.7; 211.6; 163.1; 147.9; 154.9
Steak, 29.1; ham, 20; butter, 60; milk, 306.3; bread, 482.6; potatoes, 639.8

13.

(Aggregative)		
1910–1914	1915–1919	1920–1924
100	167	128

Index numbers of Iowa farm products prices (base, 1910–1914) by geometric relative method. Other methods will give answers which approximate these.

1924	122	1931	Jan. 106		1932	Jan.	63
1925	147		Feb.	95		Feb.	57
1926	141		Mar.	98		Mar.	62
1927	140		Apr.	98		Apr.	60
1928	145		May	87		May	53
1929	147		June	82		June	49
1930	127		July	85		July	63
			Aug.	82		Aug.	61
			Sept.	78		Sept.	57
			Oct.	73		Oct.	49
			Nov.	77		Nov.	49
			Dec.	67		Dec.	43

14.

Year	V	Base-weighted		Reverse-weighted		Fisher	
		Q	P	Q	P	Q	P
(Base) 1900	100	100	100	100	100	100	100
1901	98	115	90	108.9	85.2	111.95	87.60
1902	136	145	95	143.2	93.8	144.10	94.40
1903	98	115	85	115.3	85.2	115.15	85.10
1904	117	130	90	130.0	90.0	130.00	90.00
1905	127	120	105	121.0	105.8	120.50	105.40
1906	152	130	115	132.2	116.9	131.10	115.95
1907	150	130	115	130.4	115.4	130.20	115.20
1908	192	155	125	153.6	123.9	154.30	124.45

Weights for relative method = values 100, 50, and 50 reduced to percentages 50, 25, and 25. Results check with base-weighted price index, as given above.

CHAPTER VI

TIME SERIES: TRENDS

It has been seen that an average is strictly a point in the magnitude scale of a distribution which equates the positive and negative deviations from it, or in general which indicates the central tendency of a group of items. In the discussion of index numbers, methods for obtaining successive averages of prices or other data were considered, thus extending the idea of an average to a time interval. An index number of prices thus obtained resembles, in some degree, a trend line, in that it pictures an average movement through successive years. Hence index numbers serve as a convenient point of departure for the consideration of trends.

The nature of a trend.—In fitting a trend to data, it is necessary to go a little farther than in index numbers and attempt to determine a smoothed line calculated to measure the general direction of change. Thus, in attempting to describe the rise of prices from 1914 to 1920, the first requirement is a representative index number for each year. Such index numbers would exhibit from year to year an irregular upward movement. This general upward movement might then be smoothed by drawing a straight line through it, following as closely as possible the irregular course of the data (cf. Chart 17). The line thus illustrated does not represent the so-called secular or long-time trend, but it serves as a measure of the average rate of change during the interval in question. This rate is 24.11 points a year, or $24.11 \div T$ in percentages. Or, if the irregularities moved roughly along a curved path, then a smoothed curve such as a parabola might be fitted. The process is analogous to averaging, except that it determines a line rather than a point, and measures average change rather than average position. Just as an average to be dependable should be based upon a considerable number of items, so a trend should be based upon ample, representative data. For the sake of simplicity, the illustrations that follow employ scanty data, but it is understood that the processes thus illustrated should be more broadly applied.

In computing averages, difficulty was encountered in determining what type of mean was most appropriate under given conditions. An analogous problem arises in trend fitting, but with added complications.

The most common type of trend is a straight line, yet it is evident in most cases that, if such a line is projected much beyond the limits of the data, it has little significance. For example, the line fitted to prices for the years 1914–1920, if projected only two or three years, would have no significance at all. Nevertheless, such a trend may

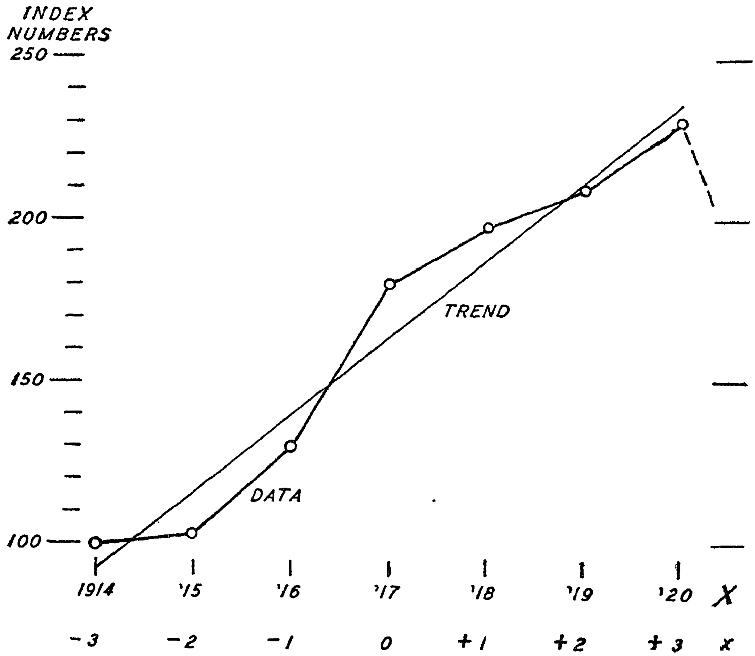


CHART 17

Index numbers of wholesale prices, United States, 1914 to 1920; base, 1914 as 100%, with straight-line trend fitted to the data. The index numbers and trend items for the successive years are: Index: 100; 103; 130; 181; 198; 210; 231. Trend: 92.4; 116.5; 140.6; 164.7; 188.8; 212.9; 237.0. Equation of trend: $T = 164.71 + 24.11x$, where x represents the years measured from the central date, 1917. The trend represents the average rise of prices during these years, but it is not the secular trend, nor could it be significantly extrapolated, that is, extended beyond 1920, since the line is fitted merely to the rising phase of the war cycle. It measures approximately the rate (dT/dx) of rise during that time, which is 24.11 points per year throughout ($dT/dx = 24.11$), or $24.11 \div T$ when expressed as a percentage ($d \log_e T/dx = \overline{dT/dx} \div \overline{164.71 + 24.11x}$). Thus the percentage rate in the middle of 1914 ($x = -3$) is $24.11 \div 92.38 = 0.261 = 26.1\%$, whereas in the middle of 1920 ($x = 3$) it is $24.11 \div 237.04 = 0.102 = 10.2\%$.

serve the purpose of averaging or measuring irregularities within the limits of the time interval to which it is applied. In the same way a regular curve may fit certain data fairly well within given limits, but be quite inappropriate if extrapolated much beyond these limits. As a rule, fluctuating trends, such as price trends, may be fitted by straight lines or parabolas, with a minimum of extrapolation. But sometimes

the type of trend is determined by the nature of the data; for example, data which tend to increase at approximately a "compound interest" rate may be fitted by a geometric or modified geometric trend, which may be plausibly extrapolated for a short period beyond the given limits. Production data over a considerable period of time are likely to approximate the so-called growth trend in one of its forms, and a short extrapolation may give a probable line of normal growth. In trend fitting some experience is necessary as a basis for the choice of a type of trend, and a most critical scrutiny of the nature of the data is necessary before any extrapolation can wisely be hazarded. After the student has become familiar with the possibilities of each type of trend, he can generally determine the type called for in a given case by an inspection of the charted data. Sometimes it may be desirable to fit a trend accurately by using all the given points in the computation, but very often approximations involving grouped data or merely a few points selected from a chart will be sufficient.

The statistical normal.*—A suitable trend fitted to adequate annual times series expresses the line of normal or average change from which cyclic and accidental deviations occur. Hence with annual data the trend may afford a basis for the computation of cyclic change. Chart 17 suggests the procedure (the ratio of data to trend), though in this case the data are incomplete, in that they form merely the ascending phase of the war cycle. Chart 18, p. 136, is a better illustration of a statistical normal from which the cycle may be calculated. However, the measurement of the cycle usually calls for more detailed figures, measuring the change by quarters, months, or weeks, and thus introducing an element of seasonal variation. Hence the application of the trend to the measurement of the cycle will be left to the next chapter.

The general method of fitting trends will first be illustrated by the straight line and parabola fitted by the so-called method of least squares. This method traces a line of the given type having the least possible

* The term "statistical normal" should not be confused with the word normal taken in the sense of justifiable or ideal. Since the statistical normal is derived from the data, it merely expresses the average course of the data artificially smoothed by the elimination of irregularities. The statistical normal of a series representing bank credit might be assumed to represent a more desirable progression than the actual, inasmuch as it would eliminate the cycle (cf. Edie, "The Banks and Prosperity"). But the statistical normal of the total volume of production might be very much below what would be considered desirable or ideal. Even in prosperity there is much wasted effort, resulting in productive inefficiency. The smoothing of a business cycle might very well eliminate extremes of price and credit inflation, but it would not represent an industrial ideal unless it lifted the volume of production to something approaching capacity.

standard deviation (the sums of the squared deviations of the data from the line a minimum) from the given points. Other less precise methods which are satisfactory for most practical purposes will also be illustrated by short examples.

Mathematical trend fitting.—In general, the strictly mathematical trends are fitted by means of trend equations which define algebraically the type of line or curve. For example, the expression

$$T = a + bx$$

defines a trend line (T) drawn on a chart through a point a which is understood to be located directly above the point of origin (the point where $x = 0$) of the horizontal (x) scale. The line is described as rising b units for each step on the x-scale. This is the straight-line trend described below. In the same way, the equation

$$T = a + bx + cx^2$$

defines the ordinary parabolic curve. The constants (parameters) a and b, or a, b, and c, may be derived from the data and substituted in the general equation. This equation is then solved for successive values of x, the time unit. For convenience in deriving the equations of the constants, the time scale is measured from some arbitrary origin, taken as $x = 0$, usually the central (average) date, and thus may have both positive and negative values. In most of the examples the year is taken as the time unit, but it is understood that the same principles may be applied to months, or other periods. The chief advantage of fitting trends by general equations is that they may readily be extrapolated; that is, the trend may be calculated for a year or more in advance of current data as an estimate of prospective change. The reliability of such an estimate, however, is a matter of judgment in the specific case.

The straight-line trend, method of least squares.—The general equation of the straight-line trend (T) fitted to successive items in a time series (Y) is

$$T = a + bx$$

where $x = 0$ at the central or average date. This equation defines the trend as one which starts with the height a on the ordinate $x = 0$, and rises b units in each successive time interval, x. The value of a and b for a given problem may be found from the data by the following equations of the constants (method of least squares; time scale regular),

$$a = \Sigma Y/n; \quad b = \Sigma xY/\Sigma x^2$$

When these values have been found, they replace a and b in the equation $T = a + bx$, and the successive trend points (T) may then be

found by substituting successive values of x in the equation. The method is illustrated in Example 38 (cf. Chart 18).

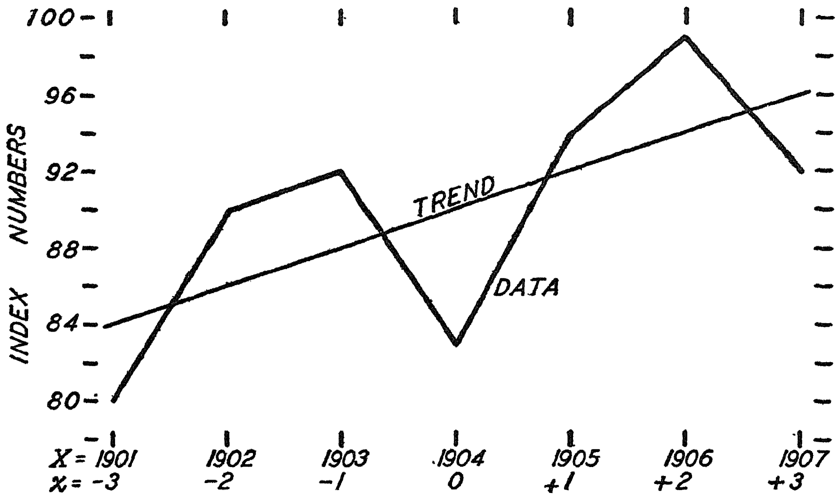


CHART 18

Straight-line trend fitted by the method of least squares to the data of Example 38. For purposes of calculation the time scale (X) is written with the central date as origin, that is, at 1904, $x = 0$. The equation of the trend is: $T = a + bx = 90 + 2x$, where $a = 90$ is the height of the trend at the point of origin (1904), and $b = 2$ is the slope of the trend per year. The rate of change is $dT/dx = 2$, and the percentage rate is $d \log_e T/dx = 2 \div \overline{90 + 2x}$, which equals 2.2% at the point of origin.

Example 38.—The straight-line trend, method of least squares. Time (X = years) remeasured from central date ($x = X - AM_x$). Trend equation, $T = a + bx$, where $a = \Sigma Y/n$ and $b = \Sigma xY/\Sigma x^2$. Trend items (T) found by substituting successive values of x in equation with computed constants, $T = 90 + 2x$; e.g., in 1901, $T = 90 + 2(-3) = 84$. The trend (T) may be more quickly written by inserting $a = 90$ at $x = 0$, and adding or subtracting $b = 2$ for the remaining items. Σx should check as zero, and ΣT should equal ΣY, except for fractional inexactness.

	X	Y	x	x^2	xY	$a + bx =$	T
	1901	80	−3	9	−240	90 − 6 =	84
	1902	90	−2	4	−180	90 − 4 =	86
	1903	92	−1	1	− 92	90 − 2 =	88
	1904	83	0	0	0	90 + 0 =	90
	1905	94	1	1	94	90 + 2 =	92
	1906	99	2	4	198	90 + 4 =	94
	1907	92	3	9	276	90 + 6 =	96
$R =$	1904	7)630	0	28	28) 56		630
	$= AM$ of years	$AM = 90$			$b = 2$		
		$= \Sigma Y/n$			$= \Sigma xY/\Sigma x^2$		

The following exceptional or irregular cases may be noted (cf. Example 39): If the trend happens to be horizontal, b will equal zero,

and the trend items will all equal a. A downward trend will register as a negative value of b. If the number of years is even, the average time is fractional, and near the center the x-series will be -1.5; -0.5; 0.5; 1.5. Or fractions may be avoided by choosing a half year as the unit, making the scale near the center: -3; -1; 1; 3; but if this is done, b is the slope for a half year. Data spaced at longer intervals, for example at each census date, may be calculated by the procedure as given, with the longer interval or decade taken as the unit. A trend may also be fitted to irregular dates by the given procedure, provided that the average time ($\Sigma X/n$) is taken as the point of origin ($x = 0$). In certain problems weights (w) may be applied to the data (Y) provided that $\Sigma wx = 0$; that is, the origin ($x = 0$) is the weighted average of the years. The weights are applied in totaling all the columns.

Example 39.—Special cases of the straight-line trend, method of least squares. (A) Origin at $1902\frac{1}{2}$, the central or average date; n an even number (4), and b negative, calling for fractional x's ($x = X - AM_x$), and bx with signs of x reversed. (B) Irregular dates, where the origin is the average year, 1910; and x is a given year less the average year, as before. A chart of this Y and T should show irregularly spaced points on a regular time scale.

A. An even number of years.

Year	Y	x	x^2	xY	$a + bx = T$
1901	94	-1.5	2.25	-141	$92 + 3.6 =$ 95.6
1902	98	-0.5	0.25	-49	$92 + 1.2 =$ 93.2
1903	86	0.5	0.25	43	$92 - 1.2 =$ 90 8
1904	90	1.5	2.25	135	$92 - 3.6 =$ 88.4
4)7610	4)368	0	5	)$-$ 12	368.0
$R = 1902.5$	$a = 92$ $= \Sigma Y/n$			$b = -2.4$ $= \Sigma xY/\Sigma x^2$	

The trend for an extrapolated year, as 1905, is

$$T = 92 - 2.4\ (1905\text{–}1902.5) = 92 - 6 = 86$$

B. Irregular dates.

Year	Y	x	x^2	xY	$a + bx = T$
1902	54	-8	64	-432	$70.5 - 16 =$ 54.5
1909	66	-1	1	-66	$70.5 - 2 =$ 68.5
1912	80	2	4	160	$70.5 + 4 =$ 74.5
1917	82	7	49	574	$70.5 + 14 =$ 84.5
4)7640	4)282	0	118	) 236	282
$R = 1910$	$a = 70.5$			$b = 2$	

The trend for any other year, as 1905, is

$$T = 70.5 + 2(1905\text{–}1910) = 70.5 - 10 = 60.5$$

The method of semi-averages, or grouped data.—A long series of data may be fitted by combining the Y's consecutively, as by threes or fives, and using the averages of the respective groups, located at their average time, in place of the Y's. Such averaging, however, usually introduces an element of inexactness. The broadest grouping, yielding a quickly computed and very useful approximation, is known as the method of semi-averages, or, in general, of grouped data. The x-scale is centered as before (each year minus the average of the years). The data (Y) are then summated in two consecutive groups of m items each (S_1 and S_2), omitting the central item if n is odd, in which case $m = (n - 1)/2$. If n is even, $m = n/2$. The equations of the constants are (method of grouped data; time scale regular):*

$$a = \Sigma Y/n$$

$$b = (S_2 - S_1) \div (m\,\overline{n - m})$$

After the constants have been obtained, the trend is computed by means of the general equation, $Y = a + bx$, as before. It will be noted that a is found as previously, but b may vary a little from that obtained by the method of least squares; e.g., for the data of Example 38 (cf. Example 40):

$$b = (S_2 - S_1) \div (m\,\overline{n - m})$$

$$= (285 - 262) \div 3 \times 4 = 1.92$$

This method need not be regarded as merely an approximation to the line of least squares. In reality it involves a different method of approach. The two summations obtained from the data in effect level the inequalities on each side of the center. Then the average deviation of the data thus leveled, considering the central item (which does not affect the slope) as if split between the two, becomes $(S_2 - S_1)/n$; and the average deviation of the x's is: $m(n - m)/n$. The slope b is the ratio of these two average deviations. The trend thus obtained is not sensitive to minor changes in the data, as is the line of least squares, but this is sometimes an advantage, as in the case of cyclical irregularities at the extremes.

Example 40.—Straight-line trend, method of semi-averages, or grouped data. General equation: $T = a + bx$, where $a = \Sigma Y/n$ and $b = (S_2 - S_1) \div (m\,\overline{n - m})$; m is the number of items on each side of the central item or point of the time series,

* If the x-scale is written at other than unit intervals, for example, if $-2\frac{1}{2}$; $-1\frac{1}{2}$; $-\frac{1}{2}$; $\frac{1}{2}$; $1\frac{1}{2}$; $2\frac{1}{2}$; is replaced by -5; -3; -1; 1; 3; 5; in order to avoid fractions (the interval from one figure to the next thus being changed to 2), or if any other interval (i) is used, the factor i must be inserted in the denominator of the formula for b. The formula for a, however, is not changed.

i.e., $m = (n - 1) \div 2$ where n is odd, and $n/2$ where n is even; S_1 is the sum of the first m items, and S_2 is the sum of the last m items. For example, if 8 items are given, S_1 and S_2 include 4 items each, and $m = 4$. The x-column becomes -3.5; -2.5; -1.5; -0.5; $+0.5$; $+1.5$; $+2.5$; $+3.5$. Whenever n is an even number the expression $(S_2 - S_1) \div (m\overline{n - m})$ may be written $4(S_2 - S_1) \div n^2$.

X	Y		x	$a + bx$	$= T$
1901	80		-3	$90 - 5.76 =$	84.24
1902	90	$S_1 = 262$	-2	$90 - 3.84 =$	86.16
1903	92		-1	$90 - 1.92 =$	88.08
1904	83		0	$90 + 0 =$	90.00
1905	94		1	$90 + 1.92 =$	91.92
1906	99	$S_2 = 285$	2	$90 + 3.84 =$	93.84
1907	92		3	$90 + 5.76 =$	95 76
	7)630				630.00
	$a = 90$				
	$= \Sigma Y/n$				

$$b = (S_2 - S_1) \div (m\overline{n - m}) = (285 - 262) \div (3\overline{7 - 3})$$
$$= 23 \div 12 = 1.92.$$

The use of semi-medians.—With very irregular data having a moderate trend, the method of semi-averages may be modified into a method of semi-medians. The items are arranged as before in two consecutive groups of m items each, where $m = n/2$ if n is even, or $(n - 1)/2$ if n is odd, in the latter case the middle item being omitted from the groups. The medians (or averages of median items) of the groups are then taken (Md_1 and Md_2), also the median of the entire series (Md). The equations of the constants then become (method of semi-medians; time scale regular).

$$a = Md$$
$$b = (Md_2 - Md_1) \div (n - m)$$

The effect is to discount the weight of the more extreme items in determining the trend.

The parabola trend, method of least squares.—If the data when charted appear to approximate a line of progressively increasing or decreasing curvature, a parabola trend (second degree or quadratic parabola) will very likely provide a suitable trend line. The general equation for this trend is

$$T = a + bx + cx^2$$

This expression is an algebraic description of a curve which crosses the point of origin at the height a, rising at that point at the rate of b per x interval and curving away from the line thus indicated by the addition of the term cx^2, which is zero at the origin. It is evident that this term

describes a curve, since the squares of x will constitute a series such as: 0; 1; 4; 9; 16; etc., the first differences of which are 1; 3; 5; 7; etc. The curve thus described by the general equation is modified by the signs and magnitudes of the constants. It can therefore be adjusted to any given set of data, with the only restriction that it must conform to the parabolic type, which implies that it has only one curvature, positive or negative, as determined by the sign of c.

In order to fit a parabola to data, the equations of the constants are required. According to the method of least squares, these equations are (regular spacings of x and $\Sigma x = 0$ are assumed):

$$b = \Sigma xY \div \Sigma x^2$$

$$c = (n\Sigma x^2Y - \Sigma x^2\Sigma Y) \div (n\Sigma x^4 - \Sigma x^2\Sigma x^2)$$

$$a = (\Sigma Y - c\Sigma x^2) \div n$$

As these equations are arranged, it is necessary to solve c before solving a. The solutions of the equations may be simplified by the use of tables, since some of the expressions are functions of n. The method is illustrated in Example 41, to which a table of certain functions of n is appended as a footnote. The data and trend are plotted in Chart 19.

Example 41.—The parabola trend, method of least squares: time (X = years) measured from central date ($x = X - AM_x$). The equations of the constants are given and solved below. Trend items (T) are found by substituting successive values of x in the general equation with computed constants, $T = 101 + 2x - 3x^2$; e.g., in 1901: $T = 101 + 2(-2) - 3(-2)^2 = 85$. For Σx^2 and $n\Sigma x^4 - \Sigma x^2\Sigma x^2$, see tables.*

* Table for computing parabola trend for given number of years, or other units (n), assuming centered time scale (x) having unit intervals (e.g., −1; 0; 1; or −1.5; −0.5; 0.5; 1.5)

n	Σx^2	$(n\Sigma x^4 - \Sigma x^2\Sigma x^2)$	n	Σx^2	$(n\Sigma x^4 - \Sigma x^2\Sigma x^2)$
2	0.5	0	14	227.5	40,768
3	2	2	15	280	61,880
4	5	16	16	340	91,432
5	10	70	17	408	131,784
6	17.5	224	18	484.5	186,048
7	28	588	19	570	257,754
8	42	1,344	20	665	351,120
9	60	2,772	21	770	471,086
10	82.5	5,280	22	885.5	623,392
11	110	9,438	23	1,012	814,660
12	143	16,016	24	1,150	1,052,480
13	182	26,026	25	1,300	1,345,500

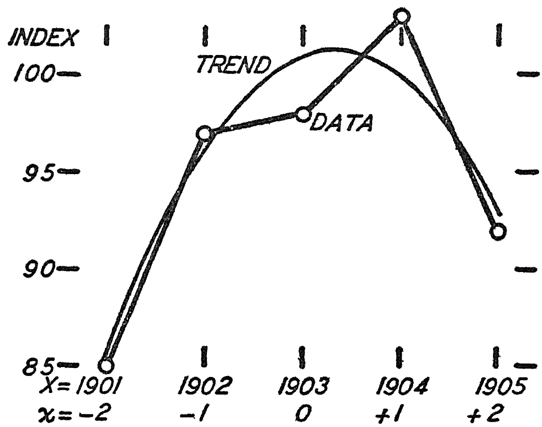


CHART 19

Parabola trend fitted to the data of Example 41 by the method of least squares. For purposes of calculation the time scale (x) is centered at the middle year, 1903, which is taken as the origin ($x = 0$). The equation of the trend is: $T = a + bx + cx^2 = 101 + 2x - 3x^2$. The mode of the parabola may be found on the x-scale as $x = -b/2c = (-2)/(-6) = 0.33$, or $1903\frac{1}{3}$ on the X-scale. The height of the curve at the mode is: $Y_{mo} = a - b^2/4c = 101 - \frac{4}{4} \div (-3) = 101.33$. The rate of change is $dT/dx = 2 - 6x$, which at the central date is the slope, $b = 2$. The percentage rate of change is $d \log_e T/dx = (dT/dx) \div T = (2 - 6x) \div (101 + 2x - 3x^2)$, which at the central date is $b/a = 2\%$. At the final date ($x = 2$) it is $(2 - 12) \div (101 + 4 - 12) = -10 \div 93 = -11\%$.

	X	Y	x	x^2	xY	x^2Y	x^4	$a + bx + cx^2 =$	T
	1901	85	−2	4	−170	340	16	101 − 4 − 12 =	85
	1902	97	−1	1	− 97	97	1	101 − 2 − 3 =	96
	1903	98	0	0	0	0	0	101 + 0 − 0 =	101
	1904	103	1	1	103	103	1	101 + 2 − 3 =	100
	1905	92	2	4	184	368	16	101 + 4 − 12 =	93
$AM =$	1903	475		10	20	908	34		475

$$b = \Sigma xY/\Sigma x^2 = 20/10 = 2$$

$$c = (n\Sigma x^2Y - \Sigma x^2\Sigma Y)/(n\Sigma x^4 - \Sigma x^2\Sigma x^2)$$

$$= (5 \times 908 - 10 \times 475)/70 = -3$$

$$a = (\Sigma Y - c\Sigma x^2)/n = [475 - (-3)\,10]/5 = 101$$

In plotting a parabola, it is often desirable to know the mode (Mo) or anti-mode (reversed mode) and its height (Y_{mo}). These may be found as: $Mo = -b/(2c)$; and $Y_{mo} = a - b^2/(4c)$, as follows:

$$Mo = -b/2c = -2/2(-3) = 0.33$$

that is, the mode of the curve is above the point $x = 0.33$ on the horizontal scale. Also

$$Y_{mo} = a - b^2/4c = 101 - 2^2/4(-3) = 101.33$$

that is, the height of the curve at the mode is $Y = 101.33$.

The above equations are found by equating the derivative of the parabola equation to zero to find the mode on the x-scale, and substituting the mode, thus expressed, in the same equation to find Y at that position.

The parabola, grouped data.—As in the case of straight-line trends, parabolas may be fitted to long series of data most conveniently by first averaging consecutively the data of some suitable interval, as three years, five years, or a decade, and then treating these averages as if they were the original data at intervals of one year.* However, the so-called method of grouped data is generally preferable to informal averaging as just described. It may be explained as follows: Select the data so that n is divisible by 3 by dropping, if necessary, one or two of the initial or end items, preferably the former. Then separate the data into three consecutive groups of an equal number of years each, indicating the number of years in each group as $m = n/3$. Total each of the three groups, indicating the totals respectively by the symbols S_1, S_2, and S_3. For example, if there were at hand annual data for the years 1901 to 1930, the first group (S_1) would include the data for 1901 to 1910, the second group (S_2) would include the data from 1911 to 1920, and the third group (S_3) would include the data from 1921 to 1930. The total number of years would be $n = 30$, and the number of years in each group would be $m = n/3 = 10$. A parabola trend may now be fitted to these groups in such a way that in each group, respectively, the sum of the trend items will equal the sum of the data. The general equation of the parabola is, as before,

$$T = a + bx + cx^2$$

and the constants b, c, and a are found by the equations,†

* For example, if index numbers were to be trended for the years 1880 to date, we might take an average for 1880 to 1884; 1885 to 1889; 1890 to 1894; etc., down through 1929. These averages then would be used as the data. The trend thus found could be charted as of 1882; 1887; 1892; etc., and extrapolated one interval at the ends. Intervening dates could be interpolated on a straight-line trend; for example, for 1883, add to the trend for 1882, one-fifth of the rise between 1882 and 1887; the next year, two-fifths of this rise, etc. The trend would be slightly inaccurate as the result of taking these averages and would consist of straight lines between five-year periods, but for most purposes it would be a workable approximation.

† If the x-scale is written at other than unit intervals, for example, if $-2\frac{1}{2}$; $-1\frac{1}{2}$; $-\frac{1}{2}$; $\frac{1}{2}$; $1\frac{1}{2}$; $2\frac{1}{2}$; is replaced by -5; -3; -1; 1; 3; 5; in order to avoid fractions the interval (i) from one figure to the next being $i = 2$, or if any other interval (i) is used, these formulas must be modified as follows:

a: no change.

b: insert factor i in denominator.

c: insert factor i^2 in denominator.

Little advantage is to be gained by this change, however.

$$b = (S_3 - S_1) \div (2m^2)$$

$$c = (S_1 + S_3 - 2S_2) \div (2m^3)$$

$$a = (S_2 - c\Sigma x_2^2) \div m$$

The process is illustrated in Example 42. The expression Σx_2^2 employed in the equation for a means the sum of the squares of the x's in the second (i.e., central) group. For example, if $n = 30$ and $m = 10$, the x-scale in the middle group would be: -4.5; -3.5; -2.5; -1.5; -0.5; 0.5; 1.5; 2.5; 3.5; 4.5; and the sum of their squares would be 82.5. The value of Σx_2^2 may be read from the table which appears as a footnote to Example 41.

Example 42.—Parabola trend fitted by method of grouped data. Data selected so that n is divisible by 3. Time scale $x = X - AM_x$. Data summated in three consecutive groups of $m = n/3$ items each: S_1, S_2, and S_3. General equation: $T = a + bx + cx^2$; constants a, b, and c, computed as indicated below, making trend equation $T = 100 + 4x - 4x^2$. Trend items then obtained by substituting successive values of x as indicated.*

X	Y	S	x	x^2	a	+	bx	+	cx^2	=	T
1901	70	S_1	−2.5	6.25	100		−10		−25	=	65
1902	80	150	−1.5	2 25	100		− 6		− 9	=	85
1903	98	S_2	−0.5	0.25	100		− 2		− 1	=	97
1904	100	198	0.5	0.25	100		+ 2		− 1	=	101
1905	95	S_3	1.5	2.25	100		+ 6		− 9	=	97
1906	87	182	2.5	6.25	100		+10		−25	=	85
		530		17.50							530

$b = (S_3 - S_1) \div 2m^2 = (182 - 150) \div (2 \times 4) = 4.$

$c = (S_1 + S_3 - 2S_2) \div 2m^3 = (150 + 182 - 396) \div (2 \times 8) = -4.$

$a = (S_2 - c\Sigma x_2^2) \div m = [198 - (-4)(0.5)] \div 2 = 100.$

* The expression Σx_2^2 means the sum of the squares of the x's in the second (i.e., central) group (S_2), the x-scale being centered ($x = 0$) at the middle of this group. The result for different values of m is $\Sigma x_2^2 = m(m^2 - 1) \div 12$ as follows (see also footnote, p. 140):

m	Σx_2^2	m	Σx_2^2	m	Σx_2^2	m	Σx_2^2
1	0	7	28.0	13	182.0	19	570.0
2	0.5	8	42.0	14	227.5	20	665.0
3	2.0	9	60.0	15	280.0	21	770.0
4	5.0	10	82.5	16	340.0	22	885.5
5	10 0	11	110.0	17	408.0	23	1012.0
6	17.5	12	143.0	18	484.5	24	1150.0

When n is not divisible by 3.—A parabola may be fitted by the grouping method to data where n is not divisible by 3, by assuming the unit of time to be one-third of that in which the data are expressed; e.g, if annual data are employed, the time unit may be taken as four months. The figure given for any year is then assumed to be repeated for each four months' interval. Thus n is made three times as large as originally given, and becomes divisible by 3. The solution proceeds as before. It is not usually necessary, however, to solve the parabola for each x, but only for every third x, choosing those falling at the middle of the year. The process is illustrated briefly in Example 43.

Example 43.—Parabola trend, method of grouped data fitted to data where n is not divisible by 3. The time unit is changed from one year to one-third year, and data are repeated for the four months' periods within each given year. A parabola trend is fitted by the method of grouped data to the data thus readjusted, where $n = 12$ and $m = 4$. The equations of the constants b, c, and a are solved below. The equation of the trend becomes $T = 102.441 + 0.625x + 0.047x^2$. The equation is solved for the middle four months' period of each year where $x = -4.5$; -1.5; 1.5; 4.5; respectively. The resulting trend items may now be taken as of the respective years. For a slight improvement on this method, readjusting the annual data at the group limits, cf. *Journal of the American Statistical Association*, September, 1927, p. 372, footnote.

Original data		Readjusted data						
Year	Index	Year	4 mo. Period	Index		x	x^2	$a + bx + cx^2 = T$
1901	100	1901	1	100		− 5.5	30.25	
			2	100		− 4.5	20.25	102.441 − 2.812 + 0.952 = 100.581
			3	100		− 3.5	12.25	
1902	103	1902	1	103	$S_1 =$ 403	− 2.5	6.25	
			2	103		− 1.5	2.25	102.441 − 0.938 + 0.106 = 101.609
			3	103		− 0.5	0.25	
1903	102	1903	1	102		0.5	0.25	
			2	102	$S_2 =$ 410	1.5	2 25	102.441 + 0.938 + 0.106 = 103.485
			3	102		2.5	6.25	
1904	107	1904	1	107		3.5	12.25	
			2	107		4.5	20.25	102.441 + 2.812 + 0.952 = 106.205
			3	107	$S_3 =$ 423	5.5	30.25	
							143.00	

$n = 12$; $m = 4$.

$b = (S_3 - S_1) \div (2m^2) = (423 - 403) \div (2 \times 4^2) = 0.625.$

$c = (S_1 + S_3 - 2S_2) \div (2m^3) = (403 + 423 - 820) \div (2 \times 4^3) = 0.047.$

$a = (S_2 - c\Sigma x_2^2) \div m = (410 - 0.047 \times 5) \div 4 = 102.441.$

The parabola, by selected points.—A convenient method of fitting a parabola with approximate accuracy consists of plotting the data (cf. Chart 20), sketching free-hand the estimated course of the trend, selecting three points on this trend as follows: Y_1, near the beginning

of the data; Y_2, near the middle of the data; and Y_3, near the end of the data. The points should be at equal distances apart on the X-scale, the time from one point to the next being represented by t. The origin of the x-scale is taken at Y_2, where $x = 0$. The height of each of the

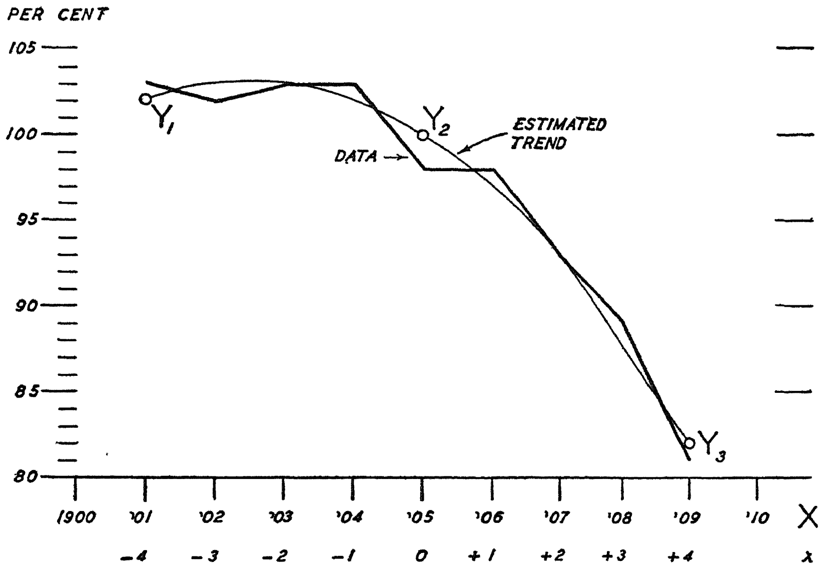


CHART 20

Parabola trend fitted by method of selected points to the data of Example 44. The data are plotted, a trend is sketched free-hand, and three points, Y_1, Y_2, and Y_3, equidistant on the x-scale, are selected on this trend as indicated. Their magnitudes 102, 100, and 82, respectively, are read from the chart. The trend is computed as indicated in Example 44.

three points is read from the chart as the magnitude of Y_1, Y_2, and Y_3, respectively. The formula of the trend may then be computed by the equations:

General equation:

$$T = a + bx + cx^2$$

Equations of the constants:

$$b = (Y_3 - Y_1) \div (2t)$$

$$c = (Y_1 + Y_3 - 2Y_2) \div (2t^2)$$

$$a = Y_2$$

The computation of the trend is illustrated in Example 44.

Example 44.—Parabola trend fitted by method of selected points. The data are charted (cf. Chart 20), a trend is sketched free-hand, three points equidistant on the

time scale are selected on this trend and their magnitudes determined by reference to the Y-scale. The equation of the trend is: $T = a + bx + cx^2 = 100 - 2.5x - 0.5x^2$; $t = 4$. If necessary, T may be centered as $T + (\Sigma Y - \Sigma T) \div n$.

Year	Y	Selected points	x	a	$+ bx$	$+ cx^2$	$= T$
1901	103	$102 = Y_1$	−4	100	+ 10	− 8	= 102
1902	102		−3	100	+ 7.5	− 4 5	= 103
1903	103		−2	100	+ 5	− 2	= 103
1904	103		−1	100	+ 2.5	− 0 5	= 102
1905	98	$100 = Y_2$	0	100	+ 0	− 0	= 100
1906	98		1	100	− 2.5	− 0 5	= 97
1907	93		2	100	− 5	− 2	= 93
1908	89		3	100	− 7.5	− 4.5	= 88
1909	81	$82 = Y_3$	4	100	− 10	− 8	= 82
	870						870

$b = (Y_3 - Y_1) \div (2t) = (82 - 102) \div (2 \times 4) = -2.5.$

$c = (Y_1 + Y_3 - 2Y_2) \div (2t^2) = (102 + 82 - 200) \div (2 \times 4^2) = -0.5.$

$a = Y_2 = 100.$

Building up parabola trends.—It is often preferable to compute parabola trends, after the equation has been found, by a process of building up, rather than by a direct use of the equation. To accomplish this, three successive points should first be computed from the equation, preferably at the center of the series (cf. Example 45). These three trend points are written in the trend column, and two additional columns are headed, respectively, first differences (Δ_1), and second differences (Δ_2). The first and second differences are found from the three computed trend points. The column of first differences may be extended backward and forward as far as desired by adding the second difference forward and subtracting it backward, since in parabolas the second difference is a constant. When the first differences have thus been obtained, the trend items themselves (T) may be found by adding the first differences forward and subtracting them backward. In both cases the process, whether adding or subtracting, is a smooth extension of the three trend figures first obtained. In using this method, care must be taken to express the first and second differences to a sufficient number of decimal places so that the error in the extreme items will not be significant.

Example 45.—Building up a parabola trend after solving the general trend equation: $T = 100 - 2.5x - 0.5x^2$. The formula is applied to the central years 1904, 1905, and 1906, where $x = -1$, 0, and $+1$, respectively. For these years T is 102, 100, and 97, respectively. The first differences are -2 and -3; that is, the trend falls two points from 1904 to 1905 and three points from 1905 to 1906. The second

difference is -1; that is, the first differences decline one point. The column Δ_1 may now be written by extending the series -2, -3 in both directions. The column T may be similarly extended by adding the first differences forward and subtracting them backwards. The work may be suitably checked by applying the formula to the extreme items, thus for 1901, $T = 100 - 2.5\ (-4) - 0.5\ (+16) = 102$.

Year	Y	x	T	Δ_1	Δ_2
1901	103	−4	102		
				1	
1902	102	−3	103		
				0	
1903	103	−2	103		
				−1	
1904	103	−1	102		
				−2	
1905	98	0	100		−1
				−3	
1906	98	1	97		
				−4	
1907	93	2	93		
				−5	
1908	89	3	88		
				−6	
1909	81	4	82		
	870		870		

The geometric trend.—The parabola trend which has just been described gives, when extended, a line which curves nearly 180 degrees. Many economic series, however, show a tendency to change their direction within a smaller compass, and sometimes to approach the horizontal line. Trends conforming to these characteristics may be best fitted by means of geometric series, either direct or modified. The geometric trend, applied directly to data, may be considered first. This trend is sometimes called the compound interest line because it resembles the line formed by charting the successive amounts of a sum of money accumulating at compound interest.

The geometric trend in its direct form is easily fitted to data by a simple modification of the straight-line formulas. To determine whether such a trend is suitable, the data may be plotted on ratio paper. If the charted figures approximate a straight line, a geometric trend is indicated. The procedure consists simply in writing the logarithms of the data (log Y) and fitting a straight-line trend to these logarithms as if they were the original figures. The method of least squares is perhaps preferable, but the method of semi-averages may be substituted as a suitable approximation in many cases. The trend thus found is obvi-

ously expressed in logarithms, hence the antilogarithms are taken as the final trend. The process is somewhat analogous to the finding of the geometric mean, which also transfers the data to the log scale and computes a result which is changed back to the original scale by taking the antilogarithm. The process is illustrated in Example 46.

Example 46.—The geometric trend. Data (Y) fitted by a trend having a geometric rate of increase (compound interest line). This trend is obtained by fitting a straight line to the logs of the data and taking the antilogs of the trend thus found. The formulas (method of least squares) $a = \Sigma Y/n$ and $b = \Sigma xY/\Sigma x^2$ are used as before but are applied to the log Y column. The time scale is centered as before.

Year	Y	log Y	x	x^2	x log Y	log T	T
1901	835	2.9217	−2	4	−5.8434	2.9211	834
1902	999	2.9996	−1	1	−2.9996	3 0003	1001
1903	1201	3.0795	0	0		3.0795	1201
1904	1439	3.1581	1	1	3.1581	3 1587	1441
1905	1732	3 2385	2	4	6.4770	3.2379	1729
		5)15 3974		10	)0.7921	15.3975	
		3.0795			0.0792		

The normal distribution curve.—The procedure that has thus been described may be applied to a parabola as well as a straight line, to yield a convex trend which, if extended symmetrically a sufficient number of terms from its mode toward its upper and lower limits, takes the form of the normal curve of distribution. Such a curve may be desirable when the data, if plotted on ratio paper, show a progressive negative curvature. In such a case a parabola trend may be fitted to the logarithms of the data, and the resulting trend transferred to the original scale by taking the antilogs. This type of trend has been found useful for trending production series (cf. Carl Snyder, " Business Cycles and Business Measurements "). Sometimes the data on ratio paper might suggest a concave curve, in which case the same procedure might be followed, but obviously the result would not conform to the curve of normal distribution. The method is not illustrated here, partly because it is simply an elaboration of the procedure illustrated in Example 46, and partly because growth trends (to be described later) will generally serve the same purpose.

The modified geometric trend.—The geometric trend fitted directly to data as described in Example 46 has rather limited uses because of its inflexibility. It can be adapted to rising or falling trends which follow the law of compound interest in direct or reverse form, but it fails to meet the requirements of a series which changes at a uniform rate after a constant has been added or subtracted, or one which rises or falls with a convex instead of a concave curve. It is therefore necessary to

consider the geometric trend adapted to a more flexible range of curves. The fitting of the modified geometric trend to each item of the data is difficult, but for most purposes the simpler method of selected points will be found satisfactory.

In order to fit a modified geometric trend to selected points, first plot the data (cf. Chart 21) and sketch, free-hand, the probable course of the trend. Then, as previously described in fitting the parabola

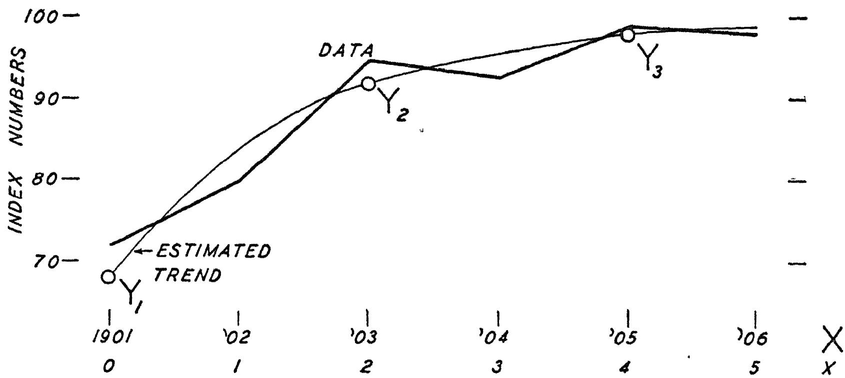


CHART 21

Modified geometric trend fitted to selected points $Y_1 = 68$, $Y_2 = 92$, $Y_3 = 98$ (cf. Example 46a). The time scale (x) has its origin at Y_1, and the time (t) between selected points is 2 years.

trend by the method of selected points (Example 44, p. 145), mark on the chart three selected equidistant T-points, one near the middle of the series and the others near the two extremes, respectively. Label the first of these points in order of time, Y_1; the second near the middle, Y_2; and the third near the end of the series, Y_3. The general equation of the curve is

$$T = a + bc^x$$

and the equations of the constants, c, b, and a, are

$$c^t = (Y_3 - Y_2) \div (Y_2 - Y_1)$$
$$b = (Y_2 - Y_1) \div (c^t - 1)$$
$$a = Y_1 - b$$

where Y_1, Y_2, and Y_3 are the selected points, as just described, and t is the number of years, or other time intervals, from Y_1 to Y_2 and from Y_2 to Y_3. These equations are based upon the assumption that the x-scale has its origin, not at the center of the series, but at the item Y_1, which may perhaps be the first item in the series, but is not necessarily so. Hence at Y_1, $x = 0$; at Y_2, $x = t$; and at Y_3, $x = 2t$.

After the equations have been solved, the trend is readily computed by writing the value of b at the date $x = 0$. The column bc^x may now be written by multiplying forward by c successively, and if necessary dividing backward. To each of the results thus obtained the constant a is added (cf. Example 46*a*; note that if $x = 0$, $c^x = 1$).

Example 46*a*.—The modified geometric trend fitted to three selected points Y_1, Y_2, and Y_3 separated at intervals of t time units (cf. Chart 21). General equation of trend $T = a + bc^x$. Time scale: at Y, $x = 0$; at Y_2, $x = t$; at Y_3, $x = 2t$. The equations of the constants are given and solved below. If T_{1900} were required, it could be found by carrying the bc^x geometric series back one term, to -64. Then, $a + bc^{-1} = 100 - 32 \div 0.5^1 = 100 - 64 = 36$. To center, take $T + (\Sigma Y - \Sigma T)/n$.

Year	Y	Selected points	x	$a + bc^x =$	T
1901	72	$68 = Y_1$	0	$100 - 32 =$	68
1902	80		1	$100 - 16 =$	84
1903	95	$92 = Y_2$	2	$100 - 8 =$	92
1904	93		3	$100 - 4 =$	96
1905	99	$98 = Y_3$	4	$100 - 2 =$	98
1906	98		5	$100 - 1 =$	99
	537	$t = 2$			537

$$c^t = (Y_3 - Y_2)/(Y_2 - Y_1) = c^2 = 6/24 = 0.25;\ c = 0.5.$$
$$b = (Y_2 - Y_1)/(c^t - 1) = 24/(-0.75) = -32.$$
$$a = Y_1 - b = 68 - (-32) = 100.$$

The modified geometric rend may be fitted by the method of grouped data, as illustrated in Example 46*b*. The data are divided into three groups of $m = n/3$ items each (S_1, S_2, and S_3); $d_1 = S_2 - S_1$, and $d_2 = S_3 - S_2$. The origin ($x = 0$) is taken at the first item. The following formulas determine the constants:

$$c^m = (S_3 - S_2) \div (S_2 - S_1) \text{ or } d_2/d_1$$
$$b = d_1(c - 1) \div (c^m - 1)^2$$
$$ma = S_1 - [d_1 \div (c^m - 1)].$$

Example 46*b*.—The modified geometric trend; method of grouped data. The data (Y) are arranged in three sub-totals (S_1, S_2, and S_3) and the first differences (d_1 and d_2) are taken. The equations of the constants as given below are then applied. General equation, $T = a + bc^x$; $m = n/3$. If n is not divisible by 3, one or two items may be dropped, or the procedure of Example 43 applied.

Year	x	Y	S	Δ_1	$100 +$	$5 \times 2^x =$	T
1900	0	104			100	5	105
1901	1	111	$S_1 = 215$		100	10	110
1902	2	122		$d_1 = 45$	100	20	120
1903	3	138	$S_2 = 260$		100	40	140
1904	4	184		$d_2 = 180$	100	80	180
1905	5	256	$S_3 = 440$		100	160	260

$$c^m = d_2/d_1;\ c^2 = 180/45 = 4;\ c = 2$$
$$b = d_1(c - 1)/(c^m - 1)^2 = 45\ (1)/(3)^2 = 5$$
$$ma = S_1 - [d_1/(c^m - 1)];\ 2a = 215 - 45/3 = 200;\ a = 100.$$

The Pearl-Reed growth curve.—The so-called logistic curve has been extensively used in trending data which represent what might be called normal growth. This trend was first popularized by Pearl and Reed in connection with certain biological and population studies. Where growth has proceeded without too great interruption, the populations and the production of the more important commodities are likely to approximate, through at least a part of their course, this type of curve (cf. Kuznets, "Secular Movements in Production and Prices"). The curve is readily fitted by selected points (modified geometric trend) as previously described, except that the method is applied to the recip-

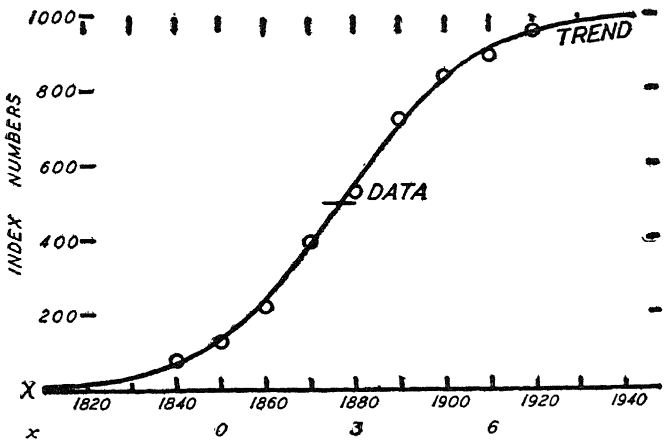


CHART 22

The Pearl-Reed growth curve fitted to the data (Y, small circles) of Example 47. The curve is fitted to the selected points $P_1 = 135$, $P_2 = 556$, $P_3 = 909$, at the dates 1850, 1880, and 1910, respectively. The time unit is a decade, and the origin ($x = 0$) is at P_1, in 1850. The trend is extrapolated at both extremes merely to indicate the shape of the completed curve, which approaches zero when extended backward, and $100{,}000/a = 1000$ when extended forward.

rocals of the data rather than to the data themselves. The trend thus found is changed back to the original scale by taking the reciprocals of each item. In finding the reciprocals of the data it is generally more convenient to write them as multiplied by a suitable power of 10, in order to avoid decimals. This will not affect the result, provided that the trend items as computed are changed to reciprocals in a like manner to give the final trend. The fitting by reciprocals is somewhat analogous to the finding of the harmonic mean, which also changes the data to a reciprocal scale. The points Y_1, Y_2, and Y_3 may be selected from a chart of the reciprocals, but it will generally be found more satisfactory to plot the data and select the three points at equidistant time intervals on the trend as graphically estimated (cf. Chart 22). The

reciprocals of these three points may then be taken as Y_1, Y_2, and Y_3, and the trend of the equation computed from them. The trend thus found will obviously be in terms of reciprocals of the required trend, hence its reciprocals are taken as the final trend. The method is illustrated in Example 47.

Example 47.—The Pearl-Reed growth curve fitted to assumed production (Y) in a certain industry; index numbers, base 1840 to 1850. Data plotted (cf. Chart 22) and three points (P) selected at equidistant time intervals on the estimated trend; the reciprocals of these points ($100{,}000/P$) are taken as Y_1, Y_2, and Y_3, respectively. General equation of trend in terms of reciprocals: $T = a + bc^x$. Equations of the constants given and solved below, making the general equation: $T = 100 + 640 \times 0.5^x$; origin 1850. The reciprocals ($100{,}000/T$) of the trend thus found are taken as the final trend.

Year	Y	x	P	$100{,}000/P$	$a + bc^x = T$	$100{,}000/T$
1840	76	-1			$100 + 1280 = 1380$	72
1850	130	0	$P_1 = 135$	$Y_1 = 740$	$100 + 640 = 740$	135
1860	220	1			$100 + 320 = 420$	238
1870	396	2			$100 + 160 = 260$	385
1880	530	3	$P_2 = 556$	$Y_2 = 180$	$100 + 80 = 180$	556
1890	725	4			$100 + 40 = 140$	714
1900	838	5			$100 + 20 = 120$	833
1910	892	6	$P_3 = 909$	$Y_3 = 110$	$100 + 10 = 110$	909
1920	959	7			$100 + 5 = 105$	952

$c^t = (Y_3 - Y_2) \div (Y_2 - Y_1) = c^3 = (-70) \div (-560) = 0.125$; $c = 0.5$.
$b = (Y_2 - Y_1) \div (c^t - 1) = (-560) \div (-0.875) = 640$.
$a = Y_1 - b = 740 - 640 = 100$.

The modified geometric trend fitted to the reciprocals does not necessarily give the Pearl-Reed curve in every case. If the data rise faster than a geometric rate, they fall outside the range of the Pearl-Reed type of curve, and the trend adjusted to them should not be described as a logistic. The logistic can easily be recognized by the fact that, if extended to its limits, it flattens out to zero at one extreme and to $1/a$ at the other extreme. It is usually a rising curve, but occasionally is reversed. When plotted it will be found to resemble rather closely the cumulatives of a normal curve of distribution; for example the binomial $(a + b)^{10}$ gives the coefficients and the cumulatives as follows:

Normal: 0, 1, 10, 45, 120, 210, 252, 210, 120, 45, 10, 1

Σ_2: 0, 1, 11, 56, 176, 386, 638, 848, 968, 1013, 1023, 1024

If these cumulatives are plotted they will resemble the extended Pearl-Reed growth curve of the positive type. In the middle portion of its

course the logistic approximates a straight line, and if the data are not sufficient to indicate the curve, the equations degenerate into those of the straight-line trend. Sometimes a case will present itself where a growth trend seems to arise from a preceding level line, rather than from zero. In this case the basic level may be subtracted from the data; or similarly a straight line with a slightly rising slope may be subtracted if this treatment seems to be required to bring the data to the logistic form. In either case the quantities subtracted at each time interval may afterwards be added to the trend items.

The moving average.—When it is necessary that the trend should closely follow the larger irregular swings of the data, the moving average may be used. From the mathematical point of view this is a very simple form of trend. It consists of averaging a certain number (m) of items (Y) beginning with the first, then with the second, etc. Suppose that the first three items are averaged: then their average is taken as the trend item at the second Y; the average of the second, third, and fourth Y's gives the trend item at the third Y, etc. Obviously, the trend will be incomplete at the two extremes, and it cannot conveniently be projected. The number of items to be included in each averaging is generally taken as approximately the length of the principal cyclic change to be eliminated; for instance, with quarterly data a four-term average will eliminate seasonal change. With an even number of items, it is theoretically desirable to include one more, giving only half weight to the two extreme items. The average will then center accurately (cf. Example 48). For example, a twelve-months' average beginning and ending with January (each January weighted one-half) will give a trend item for July. In practice, however, a twelve-month moving average centered in the seventh month is usually accurate enough. Occasionally the moving average is taken as a geometric mean if the log Y's when plotted make the secular trend more nearly a straight line. Similarly, the harmonic mean might sometimes be used. Or a trend might be removed, the moving averages of the deviations taken, and the trend again added.

One difficulty with the moving average is that it tends to "cut corners"; that is, in following an irregular cycle of the data it tends to fall within the cycle rather than to trend through the irregularities. This difficulty may be obviated by certain complex processes which will be considered in a later section of this chapter. For a full treatment of the subject the student is referred to "The Smoothing of Time Series," by Doctor F. R. Macaulay (National Bureau of Economic Research, 1931).

Example 48.—The moving average; three-term and four-term groupings, centered. The first three items of column (a) are totaled to give the first item of column (b), or 924, which is divided by 3 to give the first item of column (c), or 308. Succeeding items are similarly found by averaging the second, third, and fourth items of column (a), and so on by threes to the end of the column. In finding the centered four-year moving average (d) and (e) the totals have been doubled to avoid fractions. Thus the first item in column (d) is found by adding 296, 2 × 309, 2 × 319, 2 × 316, and 333. The total, 2517, is divided by 8 to give the first item in column (e), or 315. The remaining items are found in the same way, by moving down one item in column (a), for each calculation.

Year	(a) Index of wheat crop of world	(b) 3-year totals	(c) 3-year moving average	(d) 4-year totals (times 2) centered	(e) 4-year moving average centered
1901	296	....	....	....	....
1902	309	924	308	....	....
1903	319	944	315	2517	315
1904	316	968	323	2588	324
1905	333	992	331	2616	327
1906	343	989	330	2612	327
1907	313	974	325	2639	330
1908	318	989	330	2678	335
1909	358	1033	344	....	....
1910	357	....	....	....	....

Annual trends fitted to seasonal data.—In the analysis of time series it is commonly advisable to compute trends from annual data even though the analysis is to be made in terms of quarters, months, or weeks. Annual figures not only are more convenient in computation but also are more accurate in that a disturbing seasonal factor is eliminated. When the equation of the straight-line trend has been found on the basis of years, it is very easy to adapt it to months or other seasonal units. The term (a) in the formula is the trend height as of the center of the series. This will usually fall at January 1, or July 1. The trend item at the middle of the month following a, where $x = 1/24$, will obviously be

$$T = a + b(1/24)$$

The trend for succeeding months may be found by adding successively $b/12$, and for preceding months by subtracting successively $b/12$. For example, suppose that the trend equation for the five-year period 1911 to 1915 inclusive, has been computed from annual data as

$$T = 100 + 48x$$

and it is required to find the trend by months. The middle point of the five-year period is July 1, 1913, where $x = 0$, and the trend on that date is 100. The July, 1913, trend is ordinarily taken as of the middle of the month, $\frac{1}{24}$ of a year later, where $x = \frac{1}{24}$. Hence the trend for July, 1913, is

$$T = 100 + 48(1/24) = 102$$

The August, 1913, trend is obtained by adding to the July trend $b/12 = 4$, which gives 106. Succeeding trend items may be found by successive additions of $b/12 = 4$. Earlier trend items may be similarly obtained by subtracting $b/12 = 4$, making the June trend 98, and the May trend 94. In general, the computation thus described may be carried out on a calculating machine very quickly and accurately by carrying the addend $b/12$ to several decimal places more than are actually read. By so doing, the accumulation of error is prevented.

In the case of curve trends computed from annual data, it is generally sufficient to interpolate monthly trend points on a straight line between the annual points (taken as of July 1) by adding first $\frac{1}{24}$ of the rise of the line to the next annual point. For example, suppose that a parabola trend has been calculated from annual data representing ordinary year averages, the following measures of change have been made:

Year	x	T	Rise to next T	1/24 rise	1/12 rise
1901	−1	80	24	1	2
1902	0	104	12	0.5	1
1903	1	116			

and it is required to find trend points at intervening months. The rise of the trend line from 1901 to 1902 is 24 points (104–80); and from 1902 to 1903, 12 points (116–104). If the rise from year to year is divided by 24 it gives the average rise per one-half month of 1 and 0.5 respectively, and twice this figure, or $\frac{1}{12}$ the annual rise, is the rise per month. The monthly indexes may now easily be found beginning with $T_{1901} = 80$ which is taken as of July 1. The July trend taken as of July 15 will be 80, plus $\frac{1}{24}$ of the rise to T_{1902}; that is, plus $\frac{1}{24}$ of 24, or 1. Hence the trend for July, 1901, is 81. August may now be obtained by adding $\frac{1}{12}$ of the year's rise to the July index, or 2, making a total of $\frac{3}{24}$ of the year's rise added to 80. We now have the following trend items:

1901		1901		1902		1902	
July.......	81	Oct........	87	Jan........	93	Apr........	99
Aug........	83	Nov........	89	Feb........	95	May.......	101
Sept.......	85	Dec........	91	Mar........	97	June.......	103

In the same way the monthly trend items may be interpolated between July 1, 1902, and July 1, 1903, beginning July, 1902; 104.5; August, 1902: 105.5, etc. If an additional six items are required at the beginning and end of a series, they may easily be found by extrapolating one more item in the annual parabola and interpolating the months as before. The figures here used as illustrations, since they show an extreme rise from year to year, will not give a smooth curve, but in ordinary practice the method will provide a suitable approximation to a parabola. Of course the actual interpolation may be more precisely made by substituting in the general equation of the trend for July, 1901, the figure, $x = -23/24$; for August, 1901, $x = -21/24$, etc., and for July, 1902, $x = 1/24$, etc. If the data when charted give too irregular a line by the former method, then some such interpolation may be required. Or perhaps it would be sufficient to interpolate by means of the parabola equation, at the half unit time intervals $x = -1/2$, $x = +1/2$, etc., and then make straight-line interpolations between these half-unit and unit points by the same method as before.

SUPPLEMENTARY METHODS

The trends that are most commonly used in the statistical analysis of the social sciences have already been considered. It may be worth while, however, to give in addition the formulas for one or two other trends, and certain other methods of fitting trends already discussed. The most important additional type of trend to be considered here is the cubic parabola, or cubic. This is a parabola which, by the addition of a fourth term, provides for a secondary curvature. As a rule the cubic, if extended, does not furnish a curve which is typical of the movement of social data, and therefore it does not lend itself to extrapolation. It is sometimes useful, however, in providing a basis for a statistical normal from which to measure cycles, particularly in social data.

The formula for the cubic parabola or cubic is

$$T = a + bx + cx^2 + dx^3$$

This formula is the usual parabola with the term dx^3 added. Since x^3 takes the sign of x, it will have the effect of introducing a secondary curvature. At a considerable distance from the origin it will outweigh the other terms and will finally determine the trend. The method of fitting the cubic is so closely analogous to fitting the parabola that it will hardly be necessary to give a detailed explanation. As in the latter case, the trend may be fitted by the method of least squares. The equations are given in Table 8.

TABLE 8

Equations and tables for computing the cubic parabola, method of least squares.

I. Equations of the constants, assuming a centered time scale (x) having unit intervals (e.g., -1; 0; 1; or -1.5; -0.5; 0.5; 1.5) (cf. table for computing parabola trend).

$$c = (n\Sigma x^2Y - \Sigma x^2\Sigma Y) \div (n\Sigma x^4 - \Sigma x^2\Sigma x^2)$$

$$a = (\Sigma Y - c\Sigma x^2) \div n$$

$$d = (\Sigma x^2\Sigma x^3Y - \Sigma x^4\Sigma xY) \div (\Sigma x^2\Sigma x^6 - \Sigma x^4\Sigma x^4)$$

$$b = (\Sigma xY - d\Sigma x^4) \div \Sigma x^2$$

General equation of the cubic: $T = a + bx + cx^2 + dx^3$.

II. Table for computing the above equations of the cubic trend. For higher values of n it is preferable to combine the data consecutively, as by threes or fives fitting the trend to the averages of the groups taken at intervals of one time unit. Intermediate trend items may usually be determined by simple interpolation.

n	Σx^4	$(\Sigma x^2\Sigma x^6 - \Sigma x^4\Sigma x^4)$	n	Σx^4	$(\Sigma x^2\Sigma x^6 - \Sigma x^4\Sigma x^4)$
2	0.125	0	9	708	85,536
3	2	0	10	1208.625	254,826
4	10.25	9	11	1958	679,536
5	34	144	12	3038.75	1,656,369
6	88.375	1,134	13	4550	3,747,744
7	196	6,048	14	6608.875	7,963,956
8	388.5	24,948	15	9352	16,039,296

The cubic, like the parabola, may be fitted by special formulas to a simplified grouping of the data (cf. Table 9). The time scale (x) is centered at the middle date, as before, and the data are preferably selected so that n is divisible by 4. The data (Y) are then summated in four consecutive groups of $m = n/4$ items each (S_1, S_2, S_3, and S_4). It is also necessary to summate the second and third powers of x in the fourth group of items ($\Sigma x_4{}^2$ and $\Sigma x_4{}^3$), or to use the accompanying table. The required equations and summations are given in Table 9. If it is necessary to use this method with a series in which n is not divisible by 4, the procedure explained in connection with parabolas may be adapted (cf. Example 43, p. 144). For example, with annual data the time unit may be taken as a three months' period and the annual Y repeated for each three months. The general equation may then be solved for the magnitude of x falling at the middle of each year.

TABLE 9

Equations and tables for computing the cubic parabola, method of grouped data.

I. Equations of the constants,* assuming a centered time scale (x) having unit intervals (e.g., -1; 0; 1; or -1.5; -0.5; 0.5; 1.5; cf. table for computing parabola trend).

$$c = (S_1 + S_4 - S_2 - S_3) \div 4m^3$$

$$a = (S_1 + S_4 - 2c\Sigma x_4^2) \div 2m$$

$$d = (3S_2 + S_4 - 3S_3 - S_1) \div 6m^4$$

$$b = (S_4 - S_1 - 2d\Sigma x_4^3) \div 3m^2$$

General equation of the cubic: $T = a + bx + cx^2 + dx^3$.

II. Tables for computing the above equations of the cubic trend. Summations of powers of x in specified groups.

m	Σx_4^2	Σx_4^3	m	Σx_4^2	Σx_4^3	m	Σx_4^2	Σx_4^3
1	2.25	3.375	4	149	954	7	799.75	8985.375
2	18.5	58.5	5	291.25	2334.375	8	1194	15336
3	62.75	300.375	6	503.5	4846.5	9	1700.25	24573.375

The median cubic.—With irregular data, where the curvature is relatively small, a median cubic may be fitted by substituting m times the medians of the respective groups for the summations in the preceding equations of the grouped data method. The medians may be computed as the average of a small number of the median items.

The cubic fitted by selected points.—The cubic, like the parabola, may also be fitted to selected points. On a chart of the data, sketch the trend free-hand, and select four widely separated points, Y_1, Y_2, Y_3, and Y_4, equidistant by t units on the X-scale, and estimated to lie on the trend. The x-scale is centered midway between Y_2 and Y_3. The method follows closely that already illustrated for the parabola by selected points. The equations of the constants for the cubic are given in Table 10 (method of selected points). A trend thus computed is illustrated in Chart 23.

* If the x-scale is written at other than unit intervals, for example, if $-2\frac{1}{2}$; $-1\frac{1}{2}$; $-\frac{1}{2}$; $\frac{1}{2}$; $1\frac{1}{2}$; $2\frac{1}{2}$; is replaced by -5; -3; -1; 1; 3; 5 in order to avoid fractions (the interval (i) from one figure to the next being $i = 2$), or if any other interval (i) is used, these formulas must be modified as follows:

a: no change.
b: insert factor i in denominator.
c: insert factor i^2 in denominator.
d: insert factor i^3 in denominator.

Little advantage is to be gained by this change, however.

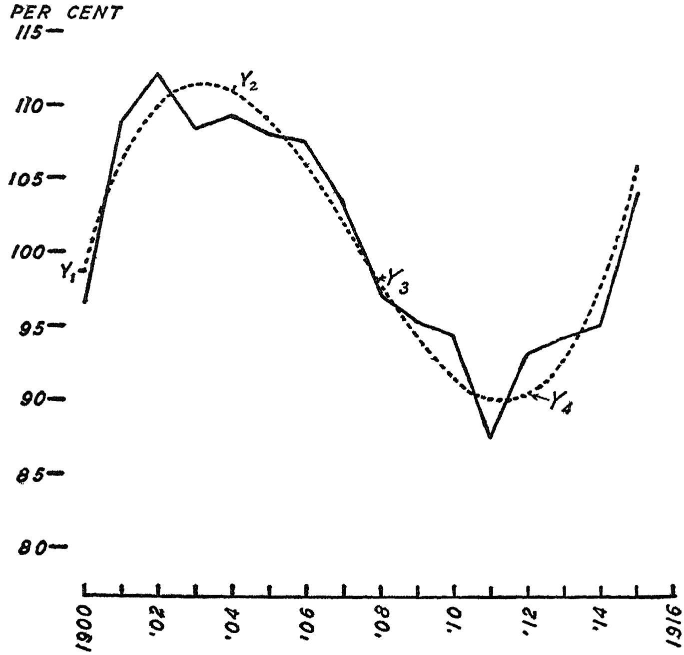


CHART 23

Cubic parabola fitted to the selected points $Y_1 = 98.5$, $Y_2 = 111.06$, $Y_3 = 98.02$, and $Y_4 = 90.1$. The equation of the trend: $T = a + bx + cx^2 + dx^3 = 105.82 - 3.58x - 0.32x^2 + 0.08x^3$. The data for the years 1900 to 1915 are: 96.40, 108.80, 112.04, 108.40, 109.23, 108.00, 107.55, 103.10, 97.00, 95.15, 94.25, 87.40, 93.20, 94.06, 95.02, 104.00. For purposes of calculation the time scale (x) is centered midway between Y_2 and Y_3 or at 1906, and the data are taken as one time unit apart.

TABLE 10

Equations for fitting the cubic parabola by the method of selected points. The method is graphically illustrated in Chart 23.

$$c = (Y_1 + Y_4 - Y_2 - Y_3) \div 4t^2$$

$$a = (Y_1 + Y_2 + Y_3 + Y_4 - 5ct^2) \div 4$$

$$d = (3Y_2 + Y_4 - Y_1 - 3Y_3) \div 6t^3$$

$$b = (Y_3 + 3Y_4 - 3Y_1 - Y_2 - 20.5dt^3) \div 10t$$

The general equation of the cubic is:

$$T = a + bx + cx^2 + dx^3$$

The summation method.—If extensive laboratory work in the fitting of parabolic trends, particularly those of higher degrees, is to be carried out, more complex methods making possible short cuts by

machine calculation are desirable. One of the best of these is the so-called summation method which has recently been elaborated by F. F. Stephan in "Summation Methods of Fitting Parabolic Curves," *Journal of the American Statistical Association*, December, 1932, pp. 413–423. It may be described and illustrated briefly as follows:

The data to which the method is applied must be a regular x-series, as of consecutive years (i.e., the ordinates must be spaced one x-unit apart). Time (or other X-unit) is centered, as in previous fittings of the quadratic and other parabolas; the origin is located at the central item or space. The data are tabulated in the usual manner consecutively, as by years (i.e., from the largest negative x to the largest positive x). The following operations are then carried out (cf. Example 49).

1. The Y column of n items is divided consecutively into two groups of m items each, as in fitting a straight-line trend by the grouped data method; that is, if n is odd, $m = (n - 1)/2$, leaving a central item, C, outside either group; and if n is even, $m = n/2$. Designate the first group of m items, A (negative x's); and the second group, B (positive x's).

2. In a column following the Y's, write a summation column (cumulative) of group A, in the usual order, i.e., beginning with the largest negative x, and working down the column. The last cumulative item thus found is designated the mth item; the next to the last, the $(m - 1)$th, and so on up the column. The whole Group A summation column thus found is designated $A1$. Similarly write a summation of group B in *reverse order;* i.e., beginning with the largest positive x, and working *up* the column. As before, designate the last cumulative item thus found (next to the x-origin) the mth item; the next to the last the $(m - 1)$th item, and so on down the column. The whole group B summation column thus found is designated $B1$.

3. Apply to the summation columns $A1$ and $B1$ the same cumulative operations just applied to groups A and B. The new summations thus obtained are written in a second column following the Y's, and are designated $A2$ and $B2$, respectively.

4. Continue to repeat operation 3 as applied to the cumulatives last found, to obtain summations $A3$ and $B3$, and so on, until the number of summations equals the number of parameters (constants) in the equation of the curve to be fitted. For example, the summations $A1$ and $B1$, $A2$ and $B2$, and $A3$ and $B3$ are sufficient for the second degree (quadratic) parabola $a + bx + cx^2$, the degree being one less than the number of parameters. The last summation column obtained may be somewhat abbreviated by omitting summation items not called for in later equations.

5. From the summation columns thus computed select the items designated below, and calculate the required values of D as indicated ($A1$, $A2$, $B1$, etc., indicate group summations, and subscripts m, $m - 1$, etc., indicate the item in the summation as described in step 2. C is the central item).

If n is odd:

$$D_1 = B1_m + A1_m + C$$
$$D_2 = B2_m - A2_m$$
$$D_3 = B3_m + A3_m + B3_{m-1} + A3_{m-1}$$
$$D_4 = B4_{m-1} - A4_{m-1}$$
$$D_5 = B5_{m-1} + A5_{m-1} + B5_{m-2} + A5_{m-2}$$
$$D_6 = B6_{m-2} - A6_{m-2}$$

If n is even

$$D_1 = B1_m + A1_m$$
$$D_2 = B2_m - A2_m + B2_{m-1} - A2_{m-1}$$
$$D_3 = B3_{m-1} + A3_{m-1}$$
$$D_4 = B4_{m-1} - A4_{m-1} + B4_{m-2} - A4_{m-2}$$
$$D_5 = B5_{m-2} + A5_{m-2}$$
$$D_6 = B6_{m-2} - A6_{m-2} + B6_{m-3} - A6_{m-3}$$

6. Find the parameters (a, b, c, etc.) of the required parabola by the formulas on page 162 (subscripts of parameters indicate the degree of the parabola to which they are applicable: K_1, K_2, etc., refer to certain additional constants given in an Appendix, p. 346, for various values of n). Certain parameters of higher degree are obtained from those of lower degree, as indicated. For purposes of comparison the equations of the lower degrees than the one required may easily be computed.

Example 49.—The summation method applied to fitting a cubic parabola. To make the problem simple and brief, a cubic has been taken as the data, Y, hence in this case the trend $T = Y$; but the method applied to irregular data would be the same. The data are divided into two consecutive groups (A and B), of $(n - 1)/2 = m$ items each, excluding a central item C, if n is odd; or of $n/2 = m$ items each if n is even. Each group is successively summated toward the center, the last item of such a summation being designated the mth; the preceding is the $(m - 1)$th, etc. The successive summations of A are designated: $A1$, $A2$, etc., and of B: $B1$, $B2$, etc.

Parameter	n is odd	n is even
a_1	$\dfrac{D_1}{n}$	$\dfrac{D_1}{n}$
$b_1 = b_2$	$\dfrac{2D_2}{K_1}$	$\dfrac{D_2}{K_1}$
$c_2 = c_3$	$\dfrac{3D_3 - K_5D_1}{K_3}$	$\dfrac{6D_3 - K_6D_1}{K_3}$
$d_3 = d_4$	$\dfrac{10D_4 - (K_5 - 2)D_2}{1{,}000{,}000K_7}$	$\dfrac{10D_4 - K_6D_2}{2{,}000{,}000K_7}$
$e_4 = e_5$	$\dfrac{140D_5 + (K_5 - 2)(K_5D_1 - 10D_3)}{1{,}000{,}000K_9}$	$\dfrac{280D_5 + (K_6 - 3)(K_6D_1 - 20D_3)}{1{,}000{,}000K_9}$
f_5	$\dfrac{504D_6 - (7K_4 - 168)D_4 + (K_5 - 2)(K_5 - 6)D_2}{1{,}000{,}000K_{10}}$	$\dfrac{504D_6 - (7K_4 - 105)D_4 + K_6(K_6 - 3)D_2}{2{,}000{,}000K_{10}}$

n is odd or even

$$a_2 = a_3 = a_1 - \frac{K_4c_2}{12}; \qquad a_4 = a_5 = a_2 + \frac{K_4}{40}\left[\frac{K_2 - 20}{14}\,e_4\right]$$

$$b_3 = b_4 = b_1 - \frac{K_2d_3}{20}; \qquad c_4 = c_5 = c_2 - \left[\frac{K_2 - 20}{14}\,e_4\right] - e_4$$

$$b_5 = b_3 + 5K_8f_5; \qquad d_5 = d_3 - \frac{(10K_4 - 60)f_5}{36}$$

The various steps in the calculation, and the formulas employed, are stated in the accompanying text. The parameter equations are on the opposite page.

Data			Successive summations of A and B, cumulated toward the x-origin					Trend
Year	x	Y	$A1$	$A2$	$A3$	$A4$		T
1901	−5	$A =$ 129	129	129	129	129		129
1902	−4	191	320	449	578	707		191
1903	−3	250	570	1019	1,597	2,304	$(m-2)$th Σ	250
1904	−2	305	875	1894	3,491	5,795	$(m-1)$th Σ	305
1905	−1	355	1230	3124	6,615	12,410	mth Σ	355
1906	0	$C =$ 399						$a =$ 399
1907	1	$B =$ 436	2375	7271	17,101	34,336	mth Σ	436
1908	2	465	1939	4896	9,830	17,235	$(m-1)$th Σ	465
1909	3	485	1474	2957	4,934	7,405	$(m-2)$th Σ	485
1910	4	495	989	1483	1,977	2,471		495
1911	5	494	494	494	494	494		494
			$B1$	$B2$	$B3$	$B4$		

D_1	D_2	D_3	D_4
2375	7271	17,101	17,235
1230	−3124	6,615	− 5,795
399		9,830	
		3,491	
4004	4147	37,037	11,440

$$a_1 = \frac{D_1}{n} = \frac{4004}{11} = 364; \quad a_2 = a_3 = a_1 - \frac{K_4 c_2}{12} = 364 - \frac{120 \times (-3.5)}{12} = 399$$

$$b_1 = b_2 = \frac{2D_2}{K_1} = \frac{2 \times 4147}{220} = 37.7;$$

$$b_3 = b_4 = b_1 - \frac{K_2 d_3}{20} = 37.7 - \frac{356 \times (-0.16\frac{2}{3})}{20} = 40.66\tfrac{2}{3}$$

$$c_2 = c_3 = \frac{3D_3 - K_5 D_1}{K_3} = \frac{3 \times 37{,}037 - 30 \times 4004}{2574} = -\frac{9009}{2574} = -3.5$$

$$d_3 = d_4 = \frac{10D_4 - (K_5 - 2)D_2}{1{,}000{,}000K_7} = \frac{10 \times 11{,}440 - 28 \times 4147}{10{,}296} = -\frac{1{,}716}{10{,}296} = 0.16\tfrac{2}{3}$$

Trend building columns, and their successive summations for:

Positive values of x

$6d =$	−1	−1	−1	−1	−1
$2c =$	−7	−8	−9	−10	−11
$b + c + d =$	37	29	20	10	−1
$a + b + c + d =$	436	465	485	495	494
$x =$	1	2	3	4	5

Negative values of x

$-6d =$	1	1	1	1	1
$2c =$	−7	−6	−5	−4	−3
$-b + c - d =$	−44	−50	−55	−59	−62
$a - b + c - d =$	355	305	250	191	129
$x =$	−1	−2	−3	−4	−5

7. When n is odd, the trend item (T) at $x = 0$ is a. Calculate T at $x = 1$ (when n is odd), or T at $x = 0.5$ (when n is even), and at succeeding positive values, of x, by the following " building up " columns (E). Higher parameters than those required are disregarded. The process is an adaptation of the " Building up " principle explained in Example 45, p. 146.

When n is odd

$$
\begin{array}{rrrrrrl}
 & & & & & & E \\
 & & & & & 120f = & \_\_\_\_\_ \\
 & & & & 24e - & 120f = & \_\_\_\_\_ \\
 & & & 6d - & 12e + & 30f = & \_\_\_\_\_ \\
 & & 2c & + & 2e & = & \_\_\_\_\_ \\
 & b + & c + & d + & e + & f = & \_\_\_\_\_ \\
a + & b + & c + & d + & e + & f = & \_\_\_\_\_ \\
 & & & & & & \text{Total} \\
 & & & & & & x = 1
\end{array}
$$

When n is even

$$
\begin{array}{rrrrrrl}
 & & & & & & E \\
 & & & & & 120f = & \_\_\_\_\_ \\
 & & & & 24e - & 180f = & \_\_\_\_\_ \\
 & & & 6d - & 24e + & 75f = & \_\_\_\_\_ \\
 & & 2c - & 3d + & 5e - & \frac{15}{2}f = & \_\_\_\_\_ \\
 & b & & + \frac{1}{4}d & & + \frac{1}{16}f = & \_\_\_\_\_ \\
a + & \frac{1}{2}b + & \frac{1}{4}c + & \frac{1}{8}d + & \frac{1}{16}e + & \frac{1}{32}f = & \text{Total} \\
 & & & & & & x = 0.5
\end{array}
$$

Successive values of the trend, T, are calculated by successive summations of the column thus obtained, the last item in each column, including the E column, being a trend item.

Trend items at negative values of x are calculated by " building up " columns obtained as those above, except that in the tabulation the signs of the b, d, and f terms are reversed (+ changed to − and − to +). Successive summations of E are again taken, and the last items in these columns, including E, are the required trend items, as before. Results may be checked by calculating one or two trend items from the trend equation, by plotting the trend against the data, by differencing (Δ_n for n degrees is a constant), and by noting that $\Sigma Y = \Sigma T$. Other checking formulas are given by Stephan in the article previously cited.

General method of fitting parabolas.—It is sometimes necessary to fit parabolas to data which are not arranged at regular time intervals. When this is the case the equations which have previously been used

will, as a rule, not be applicable, since they have been developed on the assumption that Σx, Σx^3, and $\Sigma x^5 = 0$. This is, of course, obviously true of centered time series, but is not necessarily true of irregular x intervals, even if Σx is made to equal 0; that is, if the origin is taken at the average X. Hence in fitting parabolas to such data it will generally be necessary to use the method of normal equations. The normal equations for the quadratic parabola are as follows:

$$\begin{aligned}\Sigma Y &= na + b\Sigma x + c\Sigma x^2\\ \Sigma xY &= a\Sigma x + b\Sigma x^2 + c\Sigma x^3\\ \Sigma x^2 Y &= a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4\end{aligned}$$

and for the parabola of the next higher degree the equations may be extended symmetrically by adding the new equation $\Sigma x^3 Y = a\Sigma x^3$, etc., and the column beginning $d\Sigma x^3$, and so on, to any required degree. The equations are solved algebraically for the constants by substituting the summations of x and Y. When the x-scale is centered, the summation of odd powers of x will drop out ($\Sigma x^3 = 0$) if the time scale is regular, but not necessarily when it is irregular. The derivation of the equation is given in the Appendix (cf. p. 318).

The fitting of parabolic trends by means of the normal equations is a general method applicable to parabolas of any degree, and to regular or irregular x-scales, whether centered or not, and it is the source from which the methods previously used are derived. The coefficients and other summations, such as ΣY, Σx, Σx^2, etc., are computed from the data in the way that has already been described in the fitting of the straight line and parabola trends by the method of least squares. These coefficients, as thus computed, are inserted in the equations, and the equations are then solved by any convenient algebraic process. It will be seen that there are as many equations as parameters.

The most rapid method for solving symmetrical simultaneous equations such as the foregoing normal equations is the Doolittle method. This method is illustrated in Example 49a, where it is applied to the annual index numbers of wholesale prices in the United States, 1895–1915, as given in Exercise 12 at the close of this chapter. The example does not show the computation of the coefficients, but the method of so doing is obvious. The process is entirely mechanical and self-checking, and can be readily followed by means of the directions in the heading and in the column labeled "Operations." The solution gives the values of the parameters a, b, and c, from which the trend may be computed by methods previously explained.

Example 49*a*.—Doolittle method of solving the normal equations of a second degree parabola, as applied to annual index numbers of wholesale prices, United States, 1895–1915. For data see Exercise 12, p. 181. For a parabola of the third degree, four equations and four parameters (a, b, c and d) are written in accordance with the obvious symmetrical plan, and these may be further extended for higher degrees. For the methods of finding the coefficients and other totals (Σx^2, ΣY, ΣxY, etc.), based on a centered X-scale, see Example 41, p. 140. For purposes of calculation the summations are written under their respective column headings a, b, c (add d for a cubic, etc.) and y (ΣY, ΣxY, etc.), and the equations are designated successively, I, II, and III. The checking column is begun as the sum of rows I, II, III, respectively (IV for a cubic, etc.). The coefficients enclosed in parentheses are omitted in the operations following. The successive rows in the operations are numbered consecutively (1), (2), (3), etc., and their sources are indicated by the number of the row, or the coordinate column and row locating a single number. The designation negative (Neg.) means that the designated quotients are multiplied by minus one. $T = a + bx + cx^2 = 88.24 + 1.7831x - 0.0521x^2$, as computed below.

Normal equations:

$$na + b\Sigma x + c\Sigma x^2 = \Sigma Y$$
$$a\Sigma x + b\Sigma x^2 + c\Sigma x^3 = \Sigma xY$$
$$a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 = \Sigma x^2 Y$$

Coefficients and solution:

	a	b	c	y	(check)
I	21	0	770	1,813	2,604
II	(0)	770	0	1,373	2,143
III	(770)	(0)	50,666	65,309	116,745

Operations:

I	(1)	21	0	770	1,813	2,604	
Neg. (1)/(a, 1)	(2)	−1	0	−36.6667	−86.3333	−124	check
II	(3)		770	0	1,373	2,143	
(1)×(b, 2)	(4)		0	0	0	0	
(3)+(4)	(5)		770	0	1,373	2,143	
Neg. (5)/(b, 5)	(6)		−1	0	−1.7831	−2.7831	check
III	(7)			50,666	65,309	116,745	
(1)×(c, 2)	(8)			−28,233.3333	−66,476.6667	−95,480	
(5)×(c, 6)	(9)			0	0	0	
(7)+(8)+(9)	(10)			22,432.6667	−1,167.6667	21,265	
Neg. (10)/(c, 10)	(11)			−1	0.0521	−0.9479	check
Parameters:	(12)	88.24	1.7831	−0.0521	(−1)		

(y, 12) = − 1 (Insert)
(c, 12) = (y, 12)(y, 11) = (−1)(0.0521) = − 0.0521
(b, 12) = (c, 12)(c, 6) + (y, 12)(y, 6) = (−0.0521)(0) + (−1)(−1.7831) = 1.7831
(a, 12) = (b, 12)(b, 2) + (c, 12)(c, 2) + (y, 12)(y, 2) = (1.7831)(0) + (−0.0521)(−36.6667) + (−1)(− 86.3333) = 88.24

Growth trends fitted by grouped data.—The Pearl-Reed growth trend, which was previously described as fitted by the method of selected points, may also be easily fitted by the method of grouped data. The procedure follows closely that already described for parabolas, except that the x-scale is centered at the initial Y. The reciprocals of the data, multiplied by some convenient power of 10, are

arranged in three groups of $m = n/3$ items each, the data having been selected so that n is divisible by 3. The symbols S_1, S_2, and S_3 designate the sums of the three consecutive groups, respectively. The general equation of the modified geometric trend is (cf. p. 150),

$$T = a + bc^x$$

and the equations of the constants are

$$\begin{aligned} c^m &= (S_3 - S_2)/(S_2 - S_1); \text{ or } d_2/d_1 \\ b &= (S_2 - S_1)(c - 1)/(c^m - 1)^2 \\ a &= [S_1 - (S_2 - S_1)/(c^m - 1)] \div m \end{aligned}$$

The reciprocals of the trend thus found are taken as the final trend. The method is illustrated in Example 50.

If it seems desirable to fit the trend to a series of data where n is not divisible by 3, this may be done as in the case of the parabola and cubic already described. The time unit is taken as 4 months, and the data are repeated for each 4-month interval in the year. The origin of the x-scale is the first 4-month period of the initial year, so that for the first Y, repeated, x equals 0, 1, and 2. The curve is fitted as before, and solved for x equals 1, 4, 7, etc.

Example 50.—The Pearl-Reed growth curve fitted to assumed production (Y) in a certain industry; index numbers, base 1840–1850. Modified geometric trend (T) fitted to adjusted reciprocals of data, $100{,}000/Y$. Final trend taken as $100{,}000/T$. General equation of geometric trend: $T = a + bc^x$. Equations of constants given and solved below, making final trend equation:

$$T = 10^5 \div (100 + 1280 \times 0.5^x)$$

It is usually necessary to find the mth root by logarithms as follows:

$$\log 0.125 = 2.09691 - 3; \; (2.09691 - 3) \div 3 = 0.69897 - 1$$
$$c = \text{antilog } (0.69897 - 1) = 0.5$$

It will be observed that the log of 0.125, having a characteristic of minus one, is written 2 + mantissa − 3, the latter figure being taken equal to m. This arrangement facilitates dividing by 3, or m.

Year	Y	x	$100{,}000/Y$	$a + bc^x = T$	$100{,}000/T$
1840	76	0	1316	100 + 1280 = 1380	72
1850	130	1	769	100 + 640 = 740	135
1860	220	2	455 S_1 = 2540	100 + 320 = 420	238
1870	396	3	253	100 + 160 = 260	385
1880	530	4	189	100 + 80 = 180	556
1890	725	5	138 S_2 = 580	100 + 40 = 140	714
1900	838	6	119	100 + 20 = 120	833
1910	892	7	112	100 + 10 = 110	909
1920	959	8	104 S_3 = 335	100 + 5 = 105	952

$c^m = (S_3 - S_2) \div (S_2 - S_1) = c^3 = -245 \div -1960 = 0.125;\ c = 0.5.$

$b = (S_2 - S_1)(c - 1) \div (c^m - 1)^2 = -1960(0.5 - 1) \div (0.125 - 1)^2$

$= 980 \div 0.765625 = 1280.$

$a = [S_1 - (S_2 - S_1) \div (c^m - 1)] \div m = [2540 - (-1960) \div (0.125 - 1)]$

$\div 3 = 100.$

The Gompertz curve.—A growth curve somewhat similar to the logistic, but differing from it in that it resembles the cumulative

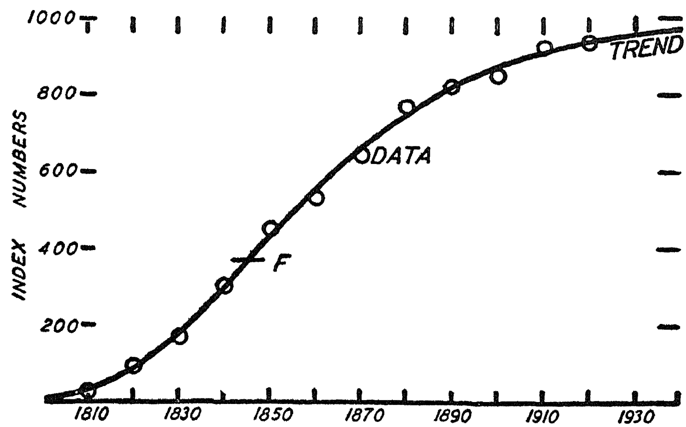


CHART 24

The Gompertz growth trend fitted by the method of grouped data to the following production index, by decades, 1810 to 1920: 29, 90, 174, 302, 450, 532, 641, 770, 820, 848, 921, 935. Base, annual averages 1910 to 1930, equation of the curve $T = ab^{cx}$. Modified geometric trend fitted to logs of data: $T = 3 - 1.525 \times 0.7^x$ ($\log a = 3$ and $\log b = 1.525$, but in solving the problem, cf. Example 51, a and b are written without log designation, since $a + bc^x$ represents a modified geometric fitted to $\log Y$), giving the trend items: 29.9, 85.6, 178.9, 299.8, 430.3, 554.2, 661.6, 748.9, 816.8, 868.0, 905.5, 932.8, and extrapolated to 1930, 952.6. These trend items are the antilogs of the trend as computed. The point of inflection of the curve (F) is located at $a/e = 1000/2.7183 = 367.9$.

of a logarithmic normal distribution rather than of a normal distribution, is known as the Gompertz curve. It may be fitted by selected points or by grouped data by first taking the logarithms of the selected points or of the data and fitting the modified geometric trend to these logarithms. The antilogs of the trend thus found are taken as the final trend (cf. Example 51). In this trend the point of inflection, where the positive curve changes from concave to convex, is at 37% of the maximum height ($1/e = 1/2.71828 = 0.36788$). This type of growth curve is obviously useful where the later stages of a series are extended beyond the symmetrical form of the

Pearl-Reed curve. More complex forms of the Gompertz curve are available in books on actuarial science.* The curve fitted by the method of grouped data is graphically illustrated in Chart 24.

Example 51.—Computation of Gompertz growth curve ($T = ab^{c^x}$) fitted to certain production statistics (Y: index numbers; base, annual averages 1810–1830). The trend is most conveniently fitted by computing the geometric trend $T = a + bc^x$ of the logs of the data. In this computation a and b are in reality logs since they are obtained from logs, but they are not designated as such. Equations of constants given and solved below, making logarithmic trend equation $T = 3 - 1.5250 \times 0.7^x$. The antilogs of the trend thus found constitute the final trend,

$$T = 1000 \times 0.02985^{\cdot 7^x}$$

					Modified geometric trend		Final trend
							Antilog
Year	x	Y	log Y		$a + bc^x$	$= T$	T
1810	0	29	1.4624		3.000 − 1.5250	1.4750	29.85
1820	1	90	1.9542		3.000 − 1 0675	1.9325	85.61
1830	2	174	2 2405		3.000 − 0.7473	2.2527	178.94
1840	3	302	2.4800	$S_1 = 8.1371$	3 000 − 0 5231	2.4769	299.85
1850	4	450	2.6532		3.000 − 0 3662	2 6338	430.33
1860	5	532	2 7259	$d_1 = 2.9354$	3.000 − 0.2563	2 7437	554.24
1870	6	641	2.8069		3.000 − 0 1794	2.8206	661.61
1880	7	770	2.8865	$S_2 = 11.0725$	3.000 − 0.1256	2 8744	748.86
1890	8	820	2.9138		3.000 − 0.0879	2 9121	816.77
1900	9	848	2 9284	$d_2 = 0.7048$	3.000 − 0 0615	2.9385	867.96
1910	10	921	2 9643		3 000 − 0.0431	2.9569	905.52
1920	11	935	2 9708	$S_3 = 11.7773$	3.000 − 0.0302	2.9698	932.82

$c^m = d_2/d_1$; $c^4 = 0.7048/2.9354 = 0.24010$; $\log c = \frac{1}{4} \log 0.24010$
$= \frac{1}{4}$ of $(3.380392 - 4) = 0.845098 - 1$; $c = \text{antilog}\ (0.845098 - 1) = 0.700$.
$b = d_1(c - 1) \div (c^m - 1)^2 = 2.9354(-0.3) \div (-0.7599)^2 = -1.5250$.
$ma = S_1 - [d_1 \div (c^m - 1)] = 8.1371 - [2.9354 \div (-0.7599)] = 8.1371 + 3.8629$
$= 12$; $a = \frac{1}{4} \times 12 = 3$.

The S-curve.—A form of the growth curve obtained graphically by the use of probability paper has been called the S-curve. It will be readily seen that if a straight line with a positive slope is drawn diagonally across probability paper, and if successive ordinates are read against the probability scale, a cumulative curve of the normal probability type is obtained. When plotted on ordinary arithmetic paper

* The Pearl-Reed and Gompertz curves may be adjusted to various degrees of asymmetry by fitting to various powers or other functions of the data. For example, if the squares, the 3/2 powers, or other functions of the data are found prior to the first step in curve fitting, the point of inflection will be modified in the final curve as fitted.

this curve is very similar to the Pearl-Reed growth curve, with the limits 0 and 100%. Hence it follows that, if data conforming in part to this type of a curve can be placed in their appropriate position on probability paper, a trend may be fitted by inspection and read at the ordinates of the data.

The difficulty in fitting such a trend lies chiefly in placing the data in their proper position on the probability curve. To do this it is necessary, first, to estimate the probable upper limit of the curve, which may be equated to 100, and the whole curve changed in the same ratio. For example, the series 1, 5, 15, 30, 42.5, 48.2, for successive years, plotted on ordinary arithmetic paper might suggest an upper limit of 50. To make the upper limit 100 would require multiplication of the series by the ratio $100/50 = 2$, giving the series, 2, 10, 30, 60, 85, 96.4. The series thus adjusted, when plotted on probability paper, may be tested by the degree to which it approximates a straight line. If some other estimate of the upper limit will give a closer approximation to a straight line, then this upper limit is to be preferred.

After the upper limit has thus been estimated, a straight-line trend may be drawn by inspection, its magnitudes read from the probability scale, and reduced to the scale of the original series by dividing by the ratio $100/50 = 2$. This trend may be further centered by adding to each trend item a correction (c) consisting of the average of the deviations of the data (Y) from the trend (T), that is, add to each T, $c = \Sigma(Y - T)/n$. The trend may be calculated more accurately by reference to tables of the normal curve which give x/σ magnitudes for each point of the data (Y). A straight-line trend may be fitted numerically to these x/σ points and read back to the corresponding area points. But this refinement of the method is hardly justified in view of the fact that the whole process is at best one of rough approximation, as determined by the method employed in placing the data upon the chart. Hence the adjusted trend fitted by inspection may be taken as appropriate for this method and may be plotted with the data on ordinary arithmetic paper as a fitted S-curve.

Other methods of estimating the upper limit of the curve are available. If it is assumed that the S-curve is closely approximated by a Pearl-Reed curve, the upper limit may be determined by use of the method of three selected points, as previously explained. For example, suppose the points 5, 30, and 48.2 are selected. Their reciprocals (times 1,000,000) are 200,000, 33,333, and 20,747, respectively; and for the purpose of this problem they may be considered as merely one time unit apart. By subtraction, $d_1 = -166{,}667$, and $d_2 = -12{,}586$. Consequently, $c = -12{,}586/-166{,}667 = 0.075516$; and $b = d_1/(c^t - 1)$

$= -166{,}667/-0.924484 = 180{,}281$. Then $a = Y_1 - b = 200{,}000 - 180{,}281 = 19{,}719$. Since the bc^x series approaches zero the curve approaches the limit a. The adjusted reciprocal of a is 1,000,000/19,719, which is 50.71, an estimate of the upper limit slightly higher than that suggested previously. Other methods of determining the upper limit might be chosen, as by fitting a parabola to the first differences of the data and taking the mode of this parabola as the 50% point, or point of inflection of the S-curve, but probably graphic estimates will be found just as satisfactory in practice.

Adjusting moving averages (MA or M_m).—It was noted earlier in this chapter that moving averages have a tendency to " cut corners "; that is, they tend to fall somewhat short of the full amplitude of the cycle. This difficulty may be obviated by the use of complex processes consisting of several moving averages of different sizes. Formulas for the smoothing of data by such a method have been used in actuarial science and have recently been adapted to economic and social data by Doctor F. R. Macaulay (cf. " The Smoothing of Time Series," National Bureau of Economic Research, 1931). One of the simpler formulas used by Doctor Macaulay is the following 27-term formula, to be applied to the smoothing of monthly data. It is computed as follows: (1) Take a 10 months' moving total of the data, using the successive weights: -1; 0; 0; $+1$; $+1$; $+1$; $+1$; 0; 0; -1. Then (2) take a 12 months' moving total of the results, and (3) take a 7 months' moving total of these totals. (4) Divide the final moving totals by 168. This graduation will fall slightly outside the parabola $y = x^2$. Such formulas give excellent results with long series of data, but with shorter series they lose too many items at the beginning and end of the series.

A comparatively simple method, somewhat similar to those just mentioned, which will approximately adjust the moving average outward to the center of the seasonal fluctuations in quarterly or monthly series is as follows: (1) Find a centered moving average (M_1) of the data. (2) Find a similar centered moving average (M_2) of the preceding moving average. (3) Take twice each item of the first of these moving averages minus the corresponding item in the second moving average ($2M_1 - M_2$) as the adjusted moving average. The theory of this adjustment is merely that the second moving average will " cut corners " about as much as the first one did and the difference between the two will therefore indicate approximately the amount that the first moving average is out of line. Therefore the first moving average plus the difference between the two ($M_1 + M_1 - M_2 = 2M_1 - M_2$) is the adjusted moving average. The process may be roughly abbre-

viated with only one moving average by assuming that the curves over a short distance are parabolas, in which case the amount of the adjustment will be (for s as an even number): *

$$(2c/6) \times [(s^2 + 2)/s^2]$$

where s is the number of subdivisions in the year (4 for quarterly data and 12 for monthly data), and $2c$ is twice the given item in the moving average less the sum of the two items located 6 months away (one-half year back and one-half year forward). The process including both the second moving average and the approximation based on the formula, together with methods of approximating numbers at the end of the series, is given in Example 52. In each case the resulting adjusted moving average may be smoothed by taking a three- or five-term moving average of it.

Example 52.—The adjusted moving average applied to seasonal data. First method (I): Two successive annual centered moving averages are found, M_1 and M_2, and the adjusted moving average is taken as $M_1 + M_1 - M_2$. (II) Second method, by parabola formula, adds to the first moving average an adjustment (d) consisting of $d = (2c/6) \times [(s^2 + 2)/s^2]$, where $2c$ is twice any given M_1 minus the two items located 6 months before and 6 months after, and s is the number of subdivisions to the year. The adjustments at the ends are made on the principle that if the moving average, or the period on which c is based, is shortened in the ratio r, the adjustment addend is multiplied by r^2, as indicated below. The data (Y) are interest rates by quarters, 1909–1913. The final results may be smoothed by a 3-term moving average.

* The same principle of adjustment of the moving average to counteract the tendency to "cut corners" may be applied to any moving average, whether of seasonal data or not, by the same procedure as is here described. The only modification of the method to be noted is a change in the parabola formula when applied to a moving average of an odd number of terms (cf. Example 52, II). In this case the formula as for months in a revised-calendar 13-month year, becomes: $d = (2c/6) \times [(s + 1) \div (s - 1)]$. In applying this formula and the corresponding one for even-term moving averages, the following transformation is convenient (t, the number of terms in the moving average is written instead of s, the number of terms in the moving average for seasonal data):

Even-term M_m: $(2c/6)[(t^2 + 2) \div t^2] = [(t^2 + 2) \div (6t^2)](2c)$.

Odd-term M_m: $(2c/6)[(t + 1) \div (t - 1)] = [(t + 1) \div 6(t - 1)](2c)$.

The first form expresses the derivation better, since $2c/6$ is the integration on the basis of a parabola, while the second factor adjusts for the "steps." The second form, however, is more convenient to usé, since, as applied to any given moving average, $2c$ is readily obtained from the data as $2M_m - M_{-t/2} - M_{+t/2}$, and the rest of the formula may be written as a single numerical coefficient.

I. First method, by second moving average (M_2).

Year	Quarter	Y	1st moving average (M_1)	2nd moving average (M_2)	Adjusted moving average ($2M_1 - M_2$)
1909	1	3.8			
	2	3.9			
	3	4.2	4.46		
	4	5.5	4.69	(4.70)*	(4.65)
1910	1	4.7	4.98	4.88	5.08
	2	4.8	5 12	4.97	5.27
	3	5 6	5.02	4.92	5.12
	4	5.3	4.80	4.74	4.86
1911	1	4.1	4 48	4.50	4.46
	2	3.6	4.18	4.28	4.08
	3	4 2	4 02	4.16	3.88
	4	4 3	4.08	4.19	3.97
1912	1	3 9	4.26	4.36	4.16
	2	4.2	4.59	4.64	4.54
	3	5.1	4.98	5.96	5.00
	4	6.0	5.34	5.28	5.40
1913	1	5 3	5.64	(5.58)	(5.88)
	2	5.7	5.72		
	3	6.0			
	4	5.8			

* One moving average in the second series (M_2) is added at each end by taking a centered 2-term moving average (shortened in the ratio $r = s/2$; $=4/2$; $r^2 = 4$). Adjusted moving average is $M_1 + r^2(M_1 - M_2)$ instead of $M_1 + M_1 - M_2$ as before.

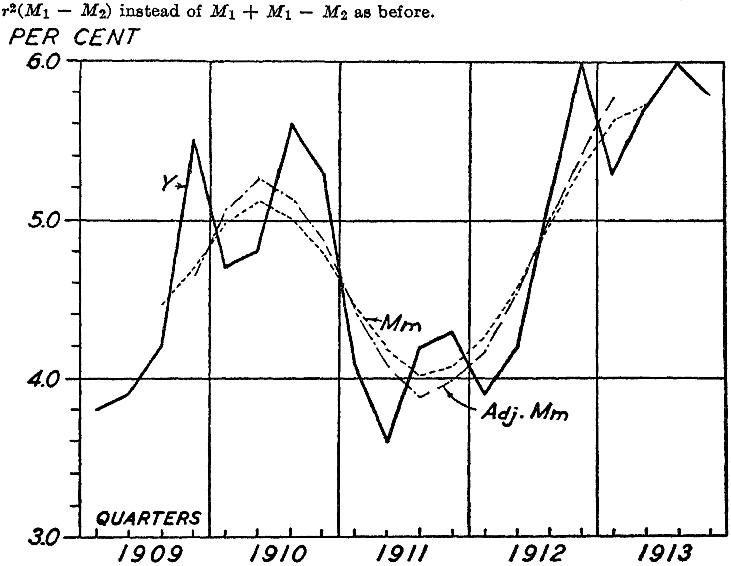


CHART 25

The adjusted moving average, revised to allow for the tendency of the moving average to "cut corners." The data are commercial interest rates in the United States, by quarters, 1909–1913. A 4-term centered moving average (M_m) is first found and is adjusted (adj. M_m) by a parabola formula as explained in Example 52, II.

II. Second method, by parabola adjustment (d) to annual centered moving average (M_m): $d = (2c/6) \times [(s^2 + 2)/s^2]$ where $2c$ is any $M_m \times 2 - M_{-s/2} - M_{+s/2}$; the subscripts $- s/2$ and $+ s/2$ indicating M_m's located 6 months preceding and 6 months following.

Year	Quarter	Y	Moving average (M_m)	$2c$	$2c \times 0.1875$	$M_m + (2c) \times 0.1875$
1909	1	3.8				
	2	3.9				
	3	4.2	4 46			
	4	5.5	4.69	(−0.24)*	(−0.04)	(4.65)
1910	1	4.7	4 98	0.48	0.09	5.07
	2	4.8	5.12	0.75	0.14	5.26
	3	5.6	5 02	0.58	0.11	5.13
	4	5.3	4.80	0 30	0.06	4.86
1911	1	4.1	4.48	−0.08	−0.02	4.46
	2	3.6	4.18	−0.52	−0.10	4.08
	3	4.2	4.02	−0.70	−0.13	3.89
	4	4.3	4.08	−0.61	−0.11	3.97
1912	1	3.9	4 26	−0.48	−0.09	4.17
	2	4 2	4 59	−0.24	−0.04	4.55
	3	5.1	4 98	0.06	0.01	4.99
	4	6.0	5.34	0.37	0.07	5 41
1913	1	5.3	5.64	(0.88)	(0.16)	(5.80)
	2	5.7	5.72			
	3	6 0				
	4	5.8				

* One extra item in the column headed $2c$ is obtained by shortening to a half year instead of a year (ratio of $r = 2$) the period from which $2c$ is obtained, and multiplying the resulting figure by $r^2 = 4$. Thus $4(2 \times 4.69 - 4.46 - 4.98) = -0.24$, which is taken as $2c$. Similarly, the last item in the $2c$ column is $4(2 \times 5.64 - 5.34 - 5.72) = 0.88$.

The sine curve.—In elaborate descriptions of cyclic change, some use has been made of the sine curve and the Fourier analysis. Such studies are beyond the scope of ordinary statistical analysis, but the sine curve may be briefly described. This curve is a wavelike line of constant amplitude and periodicity such as would be described by a marker moving forward at a constant rate and at the same time swung up and down by the rotation of a circle. There is no convenient method of fitting such a curve to data, but it may be empirically fitted by estimating first the central level or trend (a) about which it moves, its amplitude from this level or trend outward to the extremes of the swings (b), and the periodicity (p) or the number of x units from the beginning to the end of each wave. The formula for the sine curve is

$$Y = a + b \sin \theta$$

where θ is the fractional term in the expression $360x/p$. In the expression x/p, the integers are discarded because they represent merely the

number of waves from the origin. The point of origin of the X-scale as thus changed to a θ-scale is assumed to be at the point where the curve passes the central level or trend in its rising phase as indicated in Chart 26. The sines may be read from the usual table, which will give the sines from the angle $\theta = 0$ to $\theta = 90$, as indicated by the θ scale. In the second quadrant the angle is taken as $180 - \theta$. The

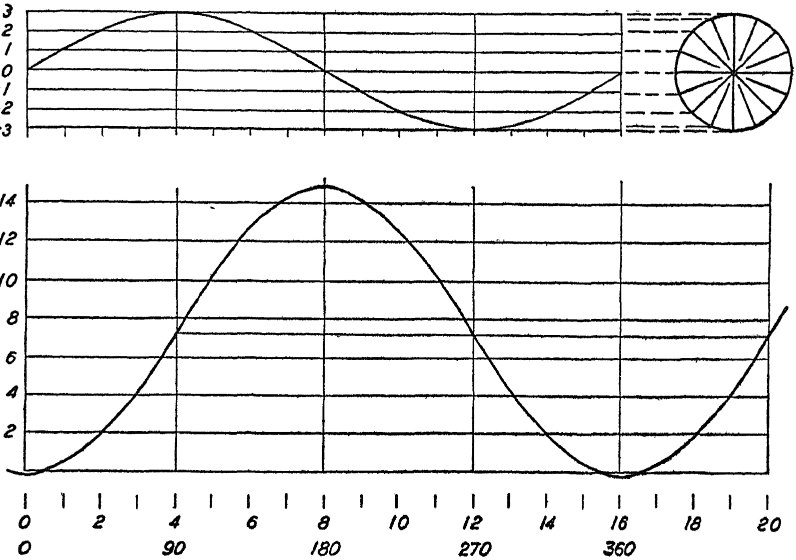


CHART 26

The sine curve. The upper figure represents a sine curve having a periodicity (p) on the x-scale (time) of 16 points and an amplitude (b) from the center level to the top or bottom of the swing of 3 units. The lower chart is similar in form, but is calculated by cumulating the upper figure. The values of Y in the upper figure are successively 0, 1, 2, 2.7, 3, 2.7, 2, 1, 0, −1, −2, −2.7, −3, −2.7, −2, −1, which cumulated give 0, 1, 3, 5.7, 8.7, 11.4, 13.4, 14.4, 14.4, 13.4, 11.4, 8.7, 5.7, 3, 1, 0, 0. The lower figure obviously lags one quadrant after the other figure. The equation of the sine curve is $Y = a + b \sin \theta$, where θ is $360x/p$. The curve as plotted is $Y = 3 \sin \theta$, and θ is $22\frac{1}{2}x$. The cumulative is here obtained only roughly without regard to the real nature of the units employed. Mathematically expressed, the cumulative or area (A) of $\sin \theta$ is $A = \int \sin \theta \times d\theta = -\cos \theta + c$ (a constant), which, adjusted to the data, becomes $A = 3\,(1 - \cos \theta)$. This would resemble the sine curve as plotted, lagged one quadrant.

third and fourth quadrants are read as $\theta - 180$ and $360 - \theta$ respectively, and the sines are given a negative sign. Thus on the basis of an estimate of the central level, the amplitude, and the periodicity, the sine curve may be written by use of a table of sines.

Trends for irregular cycles.—In measuring cycles where the data are irregular, the following three types of trends may be found useful. The first, which may be called a unit-cycle trend because the cycle is

taken as the time unit, is adapted from Wardwell (cf. Wardwell, "An Investigation of Economic Data for Major Cycles"). The second, which is formed by a process of projection, is adapted from Hall (cf. Hall, *Journal of the American Statistical Association*, June, 1926). The processes of computing these trends may be briefly described, as follows:

I. *Unit-cycle trend, cycle as time unit.*

(1) Plot data, and determine crests and troughs of cycles, drawing a vertical division line through each point (Y) marking the crests and troughs, also through the first and last items.

(2) Average each cycle from crest to crest, inclusive, as determined by the vertical division lines, giving half weight to the items (Y) at each crest, and draw a horizontal line marking the average in each cycle. Similarly average each cycle from trough to trough, and draw a horizontal line marking each average. In so doing assume the first and last items as crest or trough.

(3) In each half cycle, as determined by the vertical division lines, draw an oblique line connecting the intersections of vertical and horizontal lines. These oblique lines constitute the intermediate trend. The trend may be extended to the extremes of the data by any appropriate method of estimation, as by extending it horizontally from the closest point of intersection. Obviously the trend in the incomplete cycles at the two extremes is merely tentative.

(4) By interpolation in each half cycle, determine the ordinates of trend. If desired, these may be smoothed by means of a moving average.

II. *The projected trend.*

(1) Select some short period appropriate to the cycles to be measured, as three years. Fit a straight-line trend to the annual data of the first three years, and determine the ordinate of this trend for the third year (or for the last seasonal period of the third year if seasonal data are used). This ordinate is taken as an ordinate of the final trend.

(2) Similarly determine a trend point for the fourth year (or last seasonal period of that year) on the basis of annual data for the second, third, and fourth years. In the same way find a trend ordinate for each succeeding year to the end of the series.

(3) For trend points between the points thus determined, interpolate on a straight line. The results may be smoothed, if desired.

III. *The Fourier analysis.*

The Fourier analysis has sometimes been applied to the fitting of trends in the case of cycles as well as to the fitting of distribution curves.

It is not, however, in common use and will not be given any attention here beyond the brief statement of the method which appears in the Appendix, page 325.

Trends where time is not a coordinate.—If a straight-line trend is plotted with the data, it will be seen that the deviations $(Y - T)$ are measured vertically on the ordinates of the years; and in fitting the line it is these deviations squared that are made a minimum. This is logical with a time series, but might lead to difficulties if time is not a coordinate. Suppose, for example, that a trend were fitted to the annual prices of a given commodity plotted against the annual production. It would be a matter of debate whether the squared deviations should be made a minimum from the one axis or from the other; and a somewhat different trend would result according to the axis chosen. In such a case the deviations should preferably be a minimum when measured perpendicularly from the plotted points to the trend. The method is limited to straight-line trends. Such a fitted line may be called the intermediate straight-line trend. It is fitted as follows:

If each series (X and Y) is expressed and plotted in deviations (x and y) from its averages, and the axes are drawn at these averages as origins, then the tangent of the smaller angle (θ) formed by the trend and the axes may be found by the equation

$$\tan 2\theta = 2\Sigma xy/(\Sigma x^2 - \Sigma y^2)$$

taken as positive. The terms Σxy, Σx^2, and Σy^2 may be found by first writing each series as deviations from its average, or by the formulas

$$\Sigma x^2 = \Sigma X^2 - (\Sigma X)^2/n$$

$$\Sigma y^2 = \Sigma Y^2 - (\Sigma Y)^2/n$$

$$\Sigma xy = \Sigma XY - \Sigma X \Sigma Y/n$$

If $(\Sigma x^2 - \Sigma y^2)$ is positive, the tangent of the smaller angle as found by the above equation is the slope of the trend (b). If it is negative, the cotangent (1/tan) is the slope. In either case, b takes the sign of Σxy. The value of a is the Y average (M_y), and the trend is $T = a + bx$ in terms of the centered x-scale. In terms of the original X-scale, the trend is

$$T = (M_y - bM_x) + bX$$

This trend line has been used in measuring the elasticity of a market, on the basis of the logarithms of price and production annual link-relatives. (Cf. Schultz, "Statistical Laws of Supply and Demand.")

The derivatives of trend equations.—In connection with Chart 17, p. 133, the rate of change of the trend was indicated both as a difference and a percentage. In the case of certain trends, such as the Pearl-Reed and Gompertz curves, which approximate the cumulative or integral of the normal curve, the slopes or rates of change taken at successive points will give an approximation to a frequency distribution. Hence it is desirable to be able to express the rate of growth or derivative of a trend at any given ordinate.

The slopes of trends—or the frequencies of the distribution curves of which the trends may be taken as the cumulatives—are found by means of the derivatives of the trend equations. A derivative expresses in general terms the slope of the curve on the ordinate x, and by substituting in the derivative the specified x, the slope—or frequency—at that point on the time scale may be found. The slope of the curve on a given ordinate is defined as the slope of a tangent to the curve on that ordinate. The method of finding it for a parabola has already been discussed, and a summary of general rules, together with certain applications of the rules, including the Pearl-Reed and Gompertz derivative curves, will be found in "Mathematical Notes," pp. 316–324. The reversal of the procedure thus described gives the anti-derivative, or integral, which expresses the area under the curve from an origin to any specified ordinate.

The integration of series.—The integral or summation taken by specified increments rather than by infinitesimals, as in the case of the area under a curve, is sometimes required. For example, in the calculation of averages and trends it is often necessary to find the sum of certain regular series of numbers (e.g., $1 + 2 + 3 + \ldots n$) or their powers (e.g., $1^2 + 2^2 + 3^2 + \ldots n^2$), and for this purpose it is convenient to have at hand formulas expressing a short-cut method of obtaining these sums. Some of the more important of these formulas are given in Table 11.

TABLE 11

Formulas for obtaining the summations of certain series of integers and their powers.

I. The sum of the powers of x, when x represents the integers 1, 2, 3, . . : n.

$$\Sigma x = (n^2 + n) \div 2$$
$$\Sigma x^2 = (2n^3 + 3n^2 + n) \div 6$$
$$\Sigma x^3 = (n^4 + 2n^3 + n^2) \div 4 = (\Sigma x)^2$$
$$\Sigma x^4 = (6n^5 + 15n^4 + 10n^3 - n) \div 30$$

II. Sum of the powers of x, when x represents the integers in a centered series,

as $-3, -2, -1, 0, +1, +2, +3$, where n is an odd number representing the number of terms in the series.

$$\Sigma x = 0$$
$$\Sigma x^2 = (n^3 - n) \div 12$$
$$\Sigma x^3 = 0$$
$$\Sigma x^4 = (\Sigma x^2)(3n^2 - 7) \div 20$$

III. Sum of the powers of x, when x represents the odd integers in a centered series, as $-5, -3, -1, +1, +3, +5$, where n is an even number representing the number of terms in the series.

$$\Sigma x = 0$$
$$\Sigma x^2 = n(n + 1)(n - 1) \div 3$$
$$\Sigma x^3 = 0$$
$$\Sigma x^4 = (\Sigma x^2)(3n^2 - 7) \div 5$$

EXERCISES

1. By the method of least squares, fit straight-line trends to the following annual index numbers (consecutive years, as 1901, 1902, etc.) Plot data and trend.

(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
71	98	92	321	103	65	108	74
95	76	88	288	114	80	106	80
97	86	89	341	112	114	112	78
107	88	90	240	122	96	106	84
85	112	91	200	99	107		88
							82

(i)	(j)	(k)	(l)	(m)	(n)	(o)	(p)
90	116	104	120	109	140	119	125
99	110	106	114	111	120	123	119
97	112	101	116	106	170	126	120
106	106	92	110	97	130	124	120
111	102	94	106	99	160	124	122
109	108	85	112	90	110	123	119
					150	129	115

2. Fit straight-line trends to the following irregular series. Plot data and trend.

	(a)		(b)		(c)		(d)	(e)
Year	Index	Year	Index	Year	Index	Year	Farm wages	Index
1901	81	1920	126	1880	60	1882	18.94	103.3
1905	97	1922	114	1900	100	1885	17.97	98
1908	91	1925	120	1910	120	1888	18.24	99.5
1910	103	1929	104	1916	132	1890	18.33	100
						1892	18.60	101.5

3. Fit straight-line trends to the following indexes of American business by the method of least squares. (For current figures and other similar data see annual numbers of *Survey of Current Business.*) The base year or years are indicated by underscoring. Also find the trend item for January, 1930.

	1922	1923	1924	1925	1926	1927	1928	1929
Manufactures	87	101	94	105	108	106	111	119
Total industrial	85	101	95	104	108	106	111	118
Crop marketings	99	92	104	104	109	113	117	114
Livestock marketed	92	103	104	93	90	89	91	88
Commodity stocks	95	95	102	104	115	121	123	137
Wholesale prices	96.7	100.6	98.1	103.5	100.0	95.4	97.7	96.5
Wholesale prices, 1913 base	139	144	141	148	143	137	140	138
Price level, 1913 base	158	165	166	170	171	171	176	179
Cost of living (July, 1914 base)	173	173	174	175	174	173	171	169

Project each of the trends thus found to date and compare the results with current and recent data.

4. By the method of semi-averages, fit straight-line trends to the following annual indexes (consecutive years, as 1901, 1902, etc.) and to the data of Exercise 3. Plot data and trend.

(a)	(b)	(c)	(d)	(e)	(f)	(g)
100	90	172	104	100	95	115
101	88	170	110	92	105	126
108	80	178	109	89	100	119
		172	112	90	108	123
			120	84	112	115
					125	122

5. Compare straight-line trends fitted to the same data by the method of least squares and the method of semi-averages. What determines the difference between the two slopes? Apply both methods to these three series (consecutive time units):

(a) 100, 102, 104, 106, 108, 110
(b) 102, 102, 102, 108, 108, 108
(c) 100, 110, 120, 100, 110, 120

6. By the method of least squares, fit a straight-line trend to the following index numbers (1900–1915), find the deviations and the standard deviation, and reduce the deviations to standard deviation units (d/σ). Also tabulate the results in a frequency table of five classes. Plot the data and trend together; also plot separately the standard cycle (d/σ).

(a) 82, 97, 98, 88, 84, 85, 78, 99, 101, 115, 113, 127, 112, 114, 100, 107.
(b) 127, 118, 130, 126, 139, 123, 123, 107, 103, 80, 85, 82, 84, 92, 89, 72.

7. The following columns (Y), and the supplementary columns $8 - Y$, give all the combinations of the first seven digits having no slope. By adding a slope and a constant, any number of easily solved exercises are obtained. Thus, adding 2, 4, 6, 8, 10, 12, 14 and the constant 100 to the first column gives 105, 111, 107, 114, 112, 117, 118, the straight-line trend of which is $112 + 2x$. If quadratic parabolas are to be fitted to each column by the method of least squares, $c = (\Sigma x^2 Y - 112) \div 84$; and $a = 4 - 4c$.

3544433222211114321111322113221111113211111111
7232265775577772467766566772447766552366555444
1457724337766666545577233447663355776744666776
6775377464143227773342675655776574637675774657
2313611613624541216435747236314627265537327565
5121152151432353132624111561155432341153243233
4666546546355435654253454324532243424422432322

8. Given the following annual averages (Y_1 and Y_2), calculate the straight-line trend, by the method of least squares, for each month of the year 1913, expressing the slope of the trend to one decimal place only.

Year	Y_1	Y_2
1911	74	80
1912	72	85
1913	91	116
1914	106	123
1915	107	146

9. Fit a geometric trend (straight-line on logs) to the practice series (a) where trend should equate with data, and to the percentages of population (b) by decades, 1790–1870, living in American cities of 8000 or more.

(a) 1, 2, 4, 8, 16, 32, 64.

(b) 3.3, 4.0, 4.9, 6.7, 8.5, 12.5, 16.1, 20.9.

10. By the method of least squares, fit parabola trends to the following index numbers. Plot data and trend, and find the date at which the parabola reaches its mode, or mode inverted (maximum or minimum at $x = -b/2c$). In g and h take constants to nearest whole number. Check each trend by differencing; the second difference should be a constant.

Year	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
1901	71	97	94	120	81	91	48	51
1902	95	75	116	96	103	111	61	63
1903	97	85	106	94	103	109	72	73
1904	107	87	104	84	111	115	81	81
1905	85	111	80	106	87	89	82	81
1906	....	....	....	....	....	....	84	82
1907	....	....	....	....	....	....	76	73

11. Using the method of grouped data, fit parabola trends to the following series of annual index numbers. Check by differencing. Plot data and trend.

Year	(a)	(b)	(c)	(d)	(e)	(f)
1901	76	120	125	88	95	90
1902	83	113	121	89	97	102
1903	93	110	107	95	98	95
1904	102	95	99	102	102	105
1905	104	99	102	98	100	96
1906	95	94	96	103	92	96

12. The following series of index numbers represent wholesale prices, 1895–1915. Fit a parabola by the method of least squares, writing the constants to three decimal places. In computing the trend, solve the general equation for the three central years, take the first differences, extend them as a straight-line trend, and cumulate.

Compute also the percentage cycle. Plot together the data and trend; also plot the percentage cycle.

70, 67, 67, 70, 75, 81, 79, 84, 86, 86, 86, 89, 94, 90, 97, 101, 93, 99, 100, 98, 101

13. Using the method of grouped data, recompute the trend for the data of the preceding exercise.

14. The following index numbers represent the production of raw materials, United States, 1890–1914. Average by 5-year groups, to nearest unit, and find the constants of a parabola fitted to these averages, taken as one time-unit apart ($x = 0$ in year 1902). Compute an annual trend, to one decimal place; also the standard deviation cycle, $(Y - T)/\sigma$. Plot data and trend; also the cycle.

Year	Index	Year	Index	Year	Index	Year	Index	Year	Index
1890	53	1895	69	1900	80	1905	92	1910	100
1891	68	1896	73	1901	72	1906	98	1911	95
1892	59	1897	74	1902	89	1907	92	1912	111
1893	60	1898	79	1903	85	1908	93	1913	100
1894	58	1899	79	1904	91	1909	97	1914	108
Avg.	60	Avg.	75	Avg.	83	Avg.	94	Avg.	103

15. From the following 12 months' moving averages (Y_1 and Y_2) centered in July, extend the straight-line trend to January, 1919, using the method of semi-averages.

Year	Y_1	Y_2
1911	55	120
1912	65	130
1913	75	125
1914	70	120
1915	80	130
1916	85	145
1917	75	125
1918	65	115

16. Fit parabola trends to the following data by the method of least squares. Plot data and trend.

Year	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
1901	140	149	140	132	133	131	122	122.8
1902	120	126	160	158	120	159	116	142.8
1903	170	174	110	112	175	115	176	93.2
1904	150	154	130	134	158	139	162	114.0
1905	130	136	150	154	139	161	144	135.2
1906	110	120	170	172	118	181	122	156.8
1907	160	176	120	118	165	129	166	108.8

17. By the method of grouped data, fit parabola trends to the following series of annual index numbers. Plot data and trend. Also compute and plot the percentage cycle.

(a) 88.0, 87.5, 92.0, 100.0, 98.0, 101.0, 100.0, 104.5, 99.0.

(b) 88, 95, 93, 98, 98, 99, 105, 106, 103, 109, 111, 108, 110, 110, 113.

(c) 87, 90, 90, 92, 89; 91, 97, 94, 95, 101, 102, 102, 107, 105, 112.

18. Assume that in the preceding exercise the following three trend points for each series, respectively, have been selected, as of the middle year in each consecutive group of $n/3$ items. Compute parabola trends by the method of selected points.

(a) 89.5, 100.0, 101.5.
(b) 94.6, 104.6, 110.6.
(c) 89.4, 95.4, 105.4.

19. By the method of grouped data, fit a cubic parabola trend to the following data:

750, 908, 1000, 1162, 1040, 1178, 1254, 1340

20. By the method of grouped data, fit a cubic trend to the following annual index numbers. Plot data and trend.

(a) 96.40, 108.80, 112.04, 108.40, 109.23, 108.00, 107.55, 103.10, 97.00, 95.15, 94.25, 87.40, 93.20, 94.06, 95.02, 104.00.

(b) 58.1, 62.2, 89.4, 95.0, 95.2, 99.1, 108.0, 104.8, 99.3, 103.2, 101.9, 93.2, 94.8, 99.3, 99.7, 104.3, 106.7, 125.3, 139.0, 148.1.

21. In the preceding exercise (b), assume that the following trend points had been selected for the first, seventh, thirteenth, and nineteenth years: 54.00, 104.04, 96.48, 135.00. Compute the constants of the cubic parabola trend by the method of selected points.

22. Average the following annual index numbers by consecutive groups of threes, and fit to these averages a cubic by the method of least squares, taking the time unit as 3 years (x at 3 central years is $-\frac{1}{3}$, 0, $\frac{1}{3}$). Compute the trend for each year, and plot data and trend.

(a) 85, 80, 87, 90, 100, 98, 103, 97, 100, 104, 100, 102, 110, 107, 107.
(b) 64, 68, 66, 90, 95, 94, 98, 103, 99, 95, 100, 102, 100, 105, 101.

23. By the method of least squares fit a cubic trend to the consecutive items, 182, 196, 200, 200, 202, calculating T at x and $\frac{1}{2}x$ points. Plot data and trend. Chart also the ordinary parabola for these data.

24. Fit modified geometric trends to the following annual data. Plot data and trend.

(a) 203, 212, 210, 250, 300, 340.
(b) 100, 103, 105, 107, 118, 130.
(c) 3.4, 3.0, 2.5, 2.1, 2.12, 2.03.
(d) 90, 230, 325, 355, 370, 400.
(e) 70, 82, 90, 98, 97, 100.
(f) 740, 878, 934, 960, 990, 994, 990, 1005, 998.
(g) 20, 34, 40, 52, 43, 48, 55, 49$\frac{1}{3}$, 45.
(h) 40, 68, 80, 85, 96, 105, 110, 90.5, 97.75.
(i) 1200, 1867, 1880, 1970, 2024, 1967, 1976, 2032, 1990.

25. In Example 24, assume the following trend points for the first, third, and fifth years, and compute the constants of the trend by the method of selected points.

(a) 205; 220; 280.
(b) 101; 104; 116.
(c) 3.6; 2.4; 2.1.
(d) 80; 320; 380.
(e) 68; 92; 98.

26. Assuming that the following figures represent the population of certain cities, in thousands, according to successive censuses taken at 5-year intervals, fit a Pearl-Reed growth trend by the method of grouped data. In taking the reciprocals, divide into 100,000, writing the results as whole numbers only, except where a decimal is necessary in order to obtain three significant figures.

(a) 66.66; 87.10; 91.74; 101.80; 97.94; 101.30.
(b) 90.9; 125; 294; 476; 819; 1105.
(c) 156.2; 166.6; 188.6; 188.6; 200.0; 194.1.

27. In the preceding problem assume as trend points for the first, third, and fifth census years the following figures, and compute the constants of the trend by the method of selected points:

(a) 67.29; 94.87; 99.40.
(b) 80; 286; 800.
(c) 151.5; 185.1; 196.0.

28. Compute 3- and 5-year moving averages and a 4-year centered moving average for the following indexes of per capita physical production, 1910–1920. Plot data and both moving averages on the same chart.

100, 95, 108, 100, 98, 106, 106, 109, 108, 99, 103

29. Compute a centered annual moving average for the following commercial interest rates as given below. Also fit a straight-line trend (method of least squares) to the annual averages, and compute the trend items for each quarter, finding T for the first quarter by the equation $T = a - b(ns - 1)/2s$, and adding $b/4$ for the succeeding items consecutively ($s = 4$, or 12 with monthly data).

Quarter	1909	1910	1911	1912	1913
1st.....	3.8	4.7	4.1	3.9	5.3
2nd.....	3.9	4.8	3.6	4.2	5.7
3rd.....	4.2	5.6	4.2	5.1	6.0
4th.....	5.5	5.3	4.3	6.0	5.8
Average	4.35	5.10	4.05	4.80	5.70

30. Compute the trend items for the first and last quarters and the first and last months in the series whose annual averages are as follows:

	Year	Index		Year	Index		Year	Index
(a)	1901	92	(c)	1901	90	(e)	1901	119
	1902	96		1902	99		1902	123
	1903	94		1903	97		1903	126
	1904	108		1904	106		1904	124
	1905	110		1905	111		1905	124
				1906	109		1906	123
							1907	129
(b)	1901	94	(d)	1901	104	(f)	1901	125
	1902	98		1902	106		1902	119
	1903	100		1903	101		1903	120
	1904	110		1904	92		1904	120
	1905	106		1905	94		1905	122
				1906	85		1906	119
							1907	115

31. Interpolate linear quarterly and monthly trend items in the following parabola trends as computed from annual data:

(a)	(b)	(c)
86	64	91
94	72	82
100	78	75
104	82	70
106	84	67
	84	66

32. Fit a trend to the data for immigration, using the trend best suited to the figures. Plot both the original data and the trend as computed. For data see p. 90, *Statistical Abstract of the United States*, 1932.

33. Fit a straight-line trend to the data on Tax Yields, Personal, Total, and Corporation, Income Tax (p. 175, No. 169, *Statistical Abstract of the United States*, 1932). Plot the results obtained and the original data. To simplify the computation, the figures may be taken in millions of dollars.

34. Carl Snyder has worked out the trends for a large number of series of physical production in his book, "Business Cycles and Business Measurements." Using the data included in his study, pp. 241, 242, 243, 244, 245, and 246, compute trends on this same data and compare with the results which he obtained. Plot the results obtained from your computations on ordinary cross-section paper and compare with the results which Professor Snyder has plotted on ratio paper.

35. Compute the trends for the number of marriages and divorces, also the divorces per 1000 population. The work may be simplified by grouping the data. For data see p. 87, *Statistical Abstract of the United States*, 1932.

36. Simon S. Kuznets, in his book, "Secular Movements in Production and Prices," gives a great number of series of data which may be referred to for examples in trend fitting. The data may be worked out independently and checked against Doctor Kuznets' computations.

ANSWERS

1.

(a)	$a = 91$	$b = 4$	(h)	$a = 81$	$b = 2$
(b)	$a = 92$	$b = 4$	(i)	$a = 102$	$b = 4$
(c)	$a = 90$	$b = 0$	(j)	$a = 109$	$b = -2$
(d)	$a = 278$	$b = -29$	(k)	$a = 97$	$b = -4$
(e)	$a = 110$	$b = 0$	(l)	$a = 113$	$b = -2$
(f)	$a = 92.4$	$b = 10$	(m)	$a = 102$	$b = -4$
(g)	$a = 108$	$b = 0$	(n)	$a = 140$	$b = 0$
			(o)	$a = 124$	$b = 1$
			(p)	$a = 120$	$b = -1$

2.

(a)	$a = 93$	$b = 2$
(b)	$a = 116$	$b = -2$
(c)	$a = 103$	$b = 2$
(d)	$a = 18.416$	$b = -0.02$
(e)	$a = 100.46$	$b = -0.10$

3.

Manufacturing	$a = 103.88$	$b = 3.72$
Total industrial	$a = 103.50$	$b = 3.79$
Crop marketings	$a = 106.50$	$b = 3.12$
Livestock marketed	$a = 93.75$	$b = -1.62$
Commodity stocks	$a = 111.50$	$b = 5.98$
Wholesale prices	$a = 98.56$	$b = 0.33$

4. (a) $a = 103$: $b = 4$
(b) $a = 86$: $b = -5$
(c) $a = 173$: $b = 2$
(d) $a = 111$: $b = 3$
(e) $a = 91$: $b = -3$
(f) $a = 107.5$: $b = 5$
(g) $a = 120$: $b = 0$

	By least squares		By semi-averages	
5. (a)	$a = 105.0$	$b = 2.00$	$a = 105.0$	$b = 2.00$
(b)	$a = 105.0$	$b = 1.54$	$a = 105.0$	$b = 2.00$
(c)	$a = 110.0$	$b = 2.03$	$a = 110.0$	$b = 0$

6. (a) $a = 100$: $b = 2$: $\sigma = 10$
(b) $a = 105$: $b = -4$: $\sigma = 10$

7. $a = 4$ $b = 0$

8. T for Y_1, 1913 January 85.6; February 86.4, etc., $b = 0.8$ per month. 1913 T for Y_2 January 102.3; February 103.7, etc., $b = 1.4$ per month.

9. (a) $a = 0.9031$: $b = 0.3010$: $T = 1, 2, 4$, etc.
(b) $a = 0\ 8756$: $b = 0.1012$: $T = 2.96, 3.73, 4.71$, etc.
(c) $a = 0.69773$: $b = 0.04186$: $T = 3079, 3391$, etc.

10. (a) $a = 103$: $b = 4$: $c = -6$
(b) $a = 79$: $b = 4$: $c = 6$
(c) $a = 112$: $b = -4$: $c = -6$
(d) $a = 88$: $b = -4$: $c = 6$
(e) $a = 109$: $b = 2$: $c = -6$
(f) $a = 115$: $b = 0$: $c = -6$
(g) $a = 80$: $b = 5$: $c = -2$
(h) $a = 80$: $b = 4$: $c = -2$

11. (a) $a = 98$: $b = 5$: $c = -2$
(b) $a = 102.25$: $b = -5$: $c = 1$
(c) $a = 102.50$: $b = -6$: $c = 2$
(d) $a = 98.75$: $b = 3$: $c = -1$
(e) $a = 100\ 25$: $b = 0$: $c = -1$
(f) $a = 100.25$: $b = 0$: $c = -1$

12. $a = 88.242$: $b = 1.783$: $c = -0\ 052$

13. $a = 88.044$: $b = 1.837$: $c = -0\ 047$

14. $a = 84.286$: $b = 10.5$: $c = -0\ 643$

15. 1919 January T of $Y_1 = 81.25$; T of $Y_2 = 131.25$.

16. (a) $a = 140$: $b = 0$: $c = 0$
(b) $a = 144$: $b = 1$: $c = 1$
(c) $a = 140$: $b = 0$: $c = 0$
(d) $a = 140$: $b = 1$: $c = -1$
(e) $a = 148$: $b = 2$: $c = -1$
(f) $a = 149$: $b = 3$: $c = -1$
(g) $a = 152$: $b = 4$: $c = -2$
(h) $a = 116\ 8$: $b = 1$: $c = 2$

17. (a) $a = 100$: $b = 2$: $c = -0\ 5$
(b) $a = 104.56$: $b = 1\ 6$: $c = -0\ 08$
(c) $a = 95.44$: $b = 1.6$: $c = 0\ 08$

18. (a) $a = 100$: $b = 2$: $c = -0\ 5$
(b) $a = 104.6$: $b = 1.6$: $c = -0.08$
(c) $a = 95.4$: $b = 1.6$: $c = 0.08$

19. $a = 1000$ $b = 0$ $c = -4$ $d = 8$
$T = 708, 950, 1064, 1098, 1100, 1118, 1200, 1394$

20. (a) $a = 100$: $b = -4$: $c = 0.04$: $d = 0.08$
(b) $a = 100$: $b = -2$: $c = 0.04$: $d = 0.08$

21. $a = 100.98$: $b = -1.98$: $c = -0.08$: $d = 0.08$

22. (a) $a = 100$: $b = 2$: $c = -1$: $d = 1$
(b) $a = 100$: $b = 1$: $c = 4$: $d = 2$

23. $a = 200$: $b = 1$: $c = -2$: $d = 1$

24. (a) $a = 200$: $b = 5$: $c = 2$
(b) $a = 100$: $b = 1$: $c = 2$
(c) $a = 2$: $b = 1\ 6$: $c = 0.5$
(d) $a = 400$: $b = -320$: $c = 0\ 5$
(e) $a = 100$: $b = -32$: $c = 0\ 5$
(f) $a = 1000$: $b = -256$: $c = 0.5$
(g) $a = 50, b = -32, c = \frac{1}{2}$; $T = 18, 34, 42, 46, 48, 49, 49.5, 49.75, 49.875$.
(h) $a = 100, b = -64, c = \frac{1}{2}$; $T = 36, 68, 84, 92, 96, 98, 99, 99.5, 99.75$.
(i) $a = 2000, b = -729, c = \frac{1}{3}$; $T = 1271, 1757, 1919, 1973, 1991, 1997, 1999, 1999\frac{2}{3}, 1999\frac{8}{9}$.

25. (a) $a = 200$: $b = 5$: $c = 2$
(b) $a = 100$: $b = 1$: $c = 2$
(c) $a = 2$: $b = 1.6$: $c = 0.5$
(d) $a = 400$: $b = -320$: $c = 0.5$
(e) $a = 100$: $b = -32$: $c = 0.5$
(f) $a = 1000$: $b = -256$: $c = 0.5$

26. (a) $a = 1000$: $b = 486$: $c = \frac{1}{3}$
(b) $a = 50$: $b = 1200$: $c = 0.5$
(c) $a = 500$: $b = 160$: $c = 0.5$

27. (a) same as 23 (a)
(b) same as 23 (b)
(c) same as 23 (c)

28. 3-year moving average: 101, 101, 102, 101.3, 103.3, 107, 107.7, 105.3, 103.3.
5-year moving average: 100.2, 101.4, 103.6, 103.8, 105.4, 105.6, 105.0.
4-year moving average: 101, 102, 103, 104, 106, 106, 105.

29. Moving average

4.4625
4.6875
4.9750
5.1250
5.0250
4.8000
4.4750
4.1750
4.0250
4.0750
4.2625
4.5875
4.9750
5.3375
5.6375
5.7250

Straight-line trend: $4.80 + 0.24x$

30.

		Quarterly	Monthly
(a)	First	88.6	88.2
	Last	111.4	111.8
(b)	First	93.05	92.75
	Last	110.15	110.45
(c)	First	90.5	90.17
	Last	113.5	113.83
(d)	First	108.5	108.83
	Last	85.5	85.17
(d)	First	120.6	120.5
	Last	127.4	127.5
(f)	First	123.4	123.5
	Last	116.6	116.5

31. (a)

82.2, 84.8, (86) 87, 89, 91, 93, (94) 94.8, 96.2, 97.8, 99.2, (100) 100.5, 101.5, 102.5, 103.5, (104) 104.2, 104.8, 105.2, 105.8, (106) 106, 106.

CHAPTER VII

TIME SERIES ANALYSIS

One of the most important problems of statistical analysis is the study of time series with a view to isolating the seasonal, growth, and cyclic factors. Such a study is called time series analysis, or, colloquially, time analysis. It involves the application of principles of averaging and trending already discussed, and aims primarily at a measurement of the so-called business cycle, or similar cycles in more general social data. Such analysis is usually made on the basis of monthly data; and in the following discussion either the month or the quarter will usually be referred to as the time unit. But it is understood that the same methods of analysis may be applied to the week or other seasonal units.

An elementary analysis.—An extremely simple form of time series analysis is the comparison of data for the current month with the corresponding figure for a month ago and a year ago, or for some other earlier period taken as a base. Such comparisons will very often be found in statistical reports. For example, the following figures may be quoted from the financial papers:

Building construction contracts awarded in 37 states

June, 1931, compared with May, 1931	+ 8.4%
June, 1931, compared with June, 1930	−44.7%

Such comparisons, however, support only very limited interpretations since the month-to-month change may be largely seasonal and the year-to-year change may be due to an unmeasured growth factor. Furthermore, any such comparison may be based upon a variable or erratic item.

It is obvious that a seasonal item compared with an average of several previous years would be much more informing, provided that allowance could be made for the normal growth factor, or trend. Such a comparison may readily be made month by month or quarter by quarter. The series thus to be analyzed may be carried in two tables; the first table presents the data for several years back, and the other presents a moving total or moving average of these data. In such

a tabulation it is usually convenient to make the comparison over an even number of years, as ten. When a new monthly or quarterly item is tabulated, it may be averaged with the like months or quarters for the ten previous years to obtain a basis of comparison. This average may be called the seasonal average ($SA = \Sigma Y_s \div \overline{n+1}$). The moving total or moving average containing this new item may next be computed. The growth factor may now be readily obtained by the use of the last moving average, together with the nine preceding moving averages at the same season of the year (e.g., if the last moving average is entered at July, use the nine preceding moving averages for July). These moving averages are totaled consecutively in two groups (S_1 and S_2) of $n/2 = 5$ items each, and the growth factor (GF) is obtained as:

$$GF = 1 + 2(S_2 - S_1) \div (S_2 + S_1) \; *$$

The growth factor thus found expresses the ratio of the trend line at the end of the ten-year period as compared with the trend line at the middle of the period. The seasonal average first found ($SA = \Sigma Y_s \div \overline{n+1}$), multiplied by the growth factor (GF), will give the statistical normal, that is, the figure which might be expected on the basis of previous months or quarters assuming the rate of growth indicated by the ten years' data ending with the last item entered. If the current seasonal item is compared with this statistical normal, a percentage may be obtained indicating how this season stands relative to past experience. On the basis of this percentage it is then possible to state that

* In this formula the trend is calculated by the method of semi-averages for n years ending with the item last tabulated. The trend is expressed by the formula

$$T = a + bx = \Sigma M_m/n + x(S_2 - S_1) \div m(n - m)$$

The trend at the mid-date is a, and $n/2$ years later is $a + b(n/2)$. The change in $n/2$ years or the growth factor (GF) expressed as a ratio is

$$GF = (a + bn/2) \div a = 1 + bn/2a = 1 + n^2(S_2 - S_1) \div 2m(n - m)\Sigma M_m$$

This is the general equation for the growth factor and may be applied whether n is odd or even. But in the latter case it is readily reduced by noting that $m = n/2$, thus (n, even)

$$GF = 1 + n^2(S_2 - S_1) \div 2(n/2)^2(S_2 + S_1) = 1 + 2(S_2 - S_1) \div (S_2 + S_1)$$

In using either the general equation for the growth factor or the special equation for n as an even number, the moving totals may be used just as well as the moving averages, since in the latter case both S_1 and S_2 will be increased proportionately, and therefore the ratio $2(S_2 - S_1) \div (S_2 + S_1)$ will be unchanged.

the current item is relatively large or small. The process is illustrated in Example 53. In this example the time ($n = 4$) is too short to indicate a trend satisfactorily, but inadequate figures are used for the sake of simplifying the calculation.

As is evident from the preceding description, the form of time series analysis thus illustrated is very convenient and requires a minimum of calculation. In carrying it out currently, it is necessary only to have on hand tabulations of the data covering the required number of years, together with the annual moving totals. These may be written as two separate columns, or as two tables, by years, as in Example 53. When a new item is obtained, it may be readily evaluated by comparing it with the normal as computed from the seasonal average and the moving totals. The calculation is quickly run off on an adding machine tape. But although this method of analysis is convenient and rapid, it is not so accurate as might be desired, especially when applied to irregular data which have a marked seasonal swing.

Example 53.—Estimating the statistical normal on the basis of an even number of years (4) of seasonal data, after making allowance for trend. The normal for the first quarter of 1914 is calculated by first taking the seasonal average over 5 years, $\Sigma Y_s/(n + 1) = (4.7 + 4.1 + 3.9 + 5.3 + 4.2)/5 = 4.44$. This 5-year average is modified by the 4-year trend slope calculated from the 4 seasonal moving averages including the last one obtained. These are totaled in two consecutive groups S_1 and S_2 of $n/2$ items each, and the factor allowance for trend is calculated as $1 + 2(S_2 - S_1) \div (S_2 + S_1)$. This trend factor multiplied by the seasonal average first obtained is the statistical normal for the first quarter of 1914.

Commercial interest rates, by quarters

Quarter	1910	1911	1912	1913	1914
1	4.7	4.1	3.9	5 3	4.2
2	4.8	3.6	4.2	5.7	
3	5.6	4.2	5.1	6.0	
4	5.3	4.3	6 0	5.8	

Four-term moving average, arbitrarily centered at third term

Quarter	1910	1911	1912	1913	1914
1	.....	4.650	4.150	5.525	.....
2	.....	4.300	4.375	5.750	
3	5.100	4.050	4.800	5.700	
4	4.950	4.000	5.150	5.425	

$$N = [\Sigma Y_s/(n + 1)][1 + 2(S_2 - S_1)/(S_2 + S_1)]$$
$$= (22.2/5)[1 + 2(1.625)/(19.525)] = 5.179$$

The process as it appears on the adding machine tape, with inserted calculations, is as follows:

$$
\begin{array}{rl}
 & 4.7 \\
 & 4.1 \\
 & 3.9 \\
 & 5.3 \\
 & \underline{4.2} \\
 & 5\overline{)22.2} \text{ total} \\
\Sigma Y_s/(n+1) = & 4.44 \\
 & \\
 & 4.950 \text{ 1st } M_m \\
 & \underline{4.000} \text{ 2nd } M_m \\
S_1 = & 8.950 \text{ sub-total}
\end{array}
$$

$$
\begin{array}{rl}
 & 5.150 \text{ 3rd } M_m \\
 & \underline{5\ 425} \text{ 4th } M_m \\
S_2 + S_1 = & 19.525 \text{ sub-total} \\
 & 8.950- \\
 & \underline{8\ 950-} \\
S_2 - S_1 = & 1.625 \text{ sub-total} \\
 & \underline{1.625} \\
2(S_2 - S_1) = & 3.250 \text{ total} \\
\div (S_2 + S_1) = & 0.166453 \\
+1 = & 1.166453 \\
\times 4.44 = & 5.179
\end{array}
$$

Relative position of interest rates, first quarter of 1914: $4.2 \div 5.179 = 0.811$, or 18.9% below normal. If the moving averages have, for other reasons than appear in this calculation, been mathematically centered (cf. Example 48, p. 154), the trend may be calculated on the basis of these moving averages, but the results will differ slightly from those here given since the base period will end February 15, 1914, instead of April 1, 1914. It may be added that the moving total will serve the purpose just as well as the moving average, since $2(S_2 - S_1)/(S_2 + S_1)$ will necessarily be alike in each case. If this series were kept up currently, a new normal would be similarly computed when the second quarter, 1914, item was reported.

If it seems desirable to develop the analysis through the base period before projecting it to new items, and to determine a normal corresponding to each item of the data, this may be done on the basis of the seasonal averages and a trend based on the annual averages. Since this method, however, can be considered a rough approximation only, it will not be given any attention here, but it will be found worked out in connection with supplementary methods given later.*

The moving average percentage method.—The chief element of inaccuracy in the foregoing method of time series analysis † lies in the implied determination of the seasonal factor. Unless the data are unusually regular, medians of the items by which the seasonal is determined are preferable, so that erratic items may be discounted. But, unless the trend is nearly horizontal, medians cannot consistently be employed in finding the seasonal averages because the extreme items

* The Bureau of Business Research of the University of Iowa applied this method in 1931 to deflated bank debits for the state, covering the period of the preceding eight years. It was found that the results did not differ in any significant degree from those obtained by the more complicated methods of time series analysis commonly employed (research by Mr. T. H. Cox).

† The method is closely comparable to the time series analysis employing the method of determining the seasonal suggested by Helen D. Falkner in the *Journal of the American Statistical Association*, June, 1924. That is, the seasonal, in effect, is computed from the trend. But it has the disadvantage referred to of requiring the mean rather than the median, unless the trend slope is comparatively small.

entering into these averages are likely to be the result of the trend factor. Therefore the usual methods of analysis begin with an attempt to determine the seasonal factor on the basis of an annual moving average. Since such a moving average follows the turns of the cycle rather closely

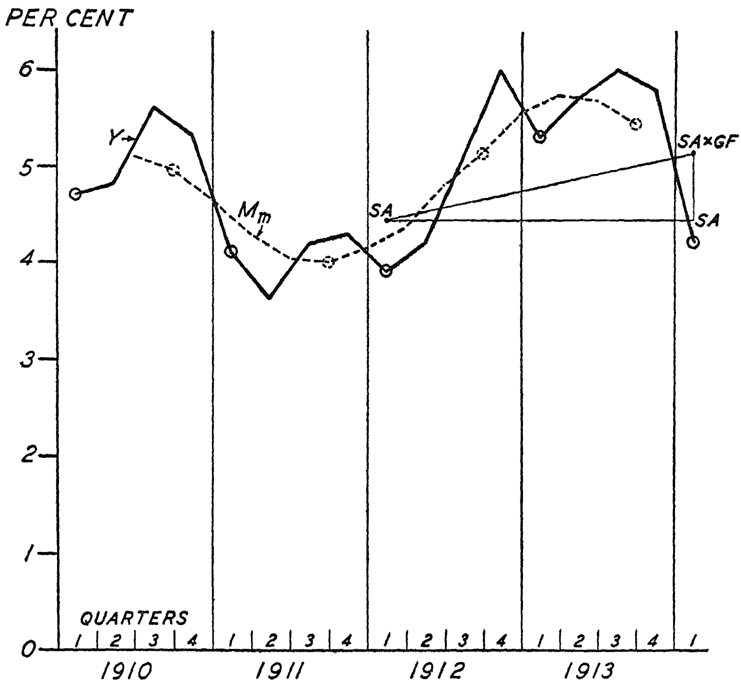


CHART 27

The statistical normal of seasonal data calculated from seasonal averages, making allowance for the growth factor, or trend. For data, see Example 53. The normal ($N = SA \times GF$) is calculated for the first quarter of 1914 on the basis of the preceding first quarters (small circles in Y line). The average of the five first quarters in the Y-line is the first quarter seasonal average (SA), which is centered in the first quarter of 1912. This seasonal average is multiplied by the growth factor, which is the ratio of the height of a trend line at the end of the four years as compared with the trend line at the middle of the four years. The annual averages from which the trend is computed are plotted as small circles in the dotted line, which is a four-term moving average. The trend is fitted to the four years ending with the first quarter of 1914.

and also conforms in general to the trend, comparisons of the data with it constitute seasonal estimates which may be combined by the median method or some modification of it. The following pages will attempt to describe the methods commonly used.

There are several methods of approach to the problem of time series analysis. With the irregular data of regional studies it is often

best to begin the analysis with the calculation and elimination of the trend, as explained later in "Supplementary Methods." But, as suggested above, the more usual methods of analysis begin with the calculation of the seasonal based upon a moving average.

Seasonal variations (S).—Seasonal variations are the fluctuations in time series due primarily to the annual variability of nature, and secondarily to custom and habit. For example, crop marketings are high in the fall and low in the spring, and the prices of farm products tend to fluctuate inversely to the marketings. This variability is found also in other series, less closely related to natural processes. For example, the Christmas trade creates a marked seasonal effect in many lines of merchandising.

Seasonal variations are difficult to measure because a time series is likely to combine trend, cyclic, and accidental variability with seasonal changes. There is therefore no absolutely accurate method of measurement. If, however, a twelve-month centered moving average is calculated from a given monthly series, the seasonal effect will be largely removed, since each item in the moving average is based upon an entire year. But it is usually desirable to go a little further and calculate the average seasonal effect, in order to have an index of seasonal influence to use in interpreting current data. Such an index may be obtained by taking the ratios (seasonal percentages, $S\%$) of the data to the moving average, month by month, and then averaging these ratios for each January, etc. The resulting average percentages tell us how far out of line from the year-to-year movement any given month may normally be expected to run. Thus we may isolate and measure one element of variability.

The process of computing the seasonal index may be described more in detail as follows: * Months are here assumed as the time units, but the same process may be applied to other intervals, as quarters or weeks.

1. Assemble and verify the data. In some cases it may be necessary to deflate by means of a suitable price index. It may also be advisable sometimes to reduce the data to an average daily figure, so as to remove irregularities arising from changes in the number of calendar or working days.

2. Find a centered twelve-month moving average, as described in the preceding chapter: the mean of thirteen months, giving half weight to extremes, or a twelve-item average arbitrarily centered in the seventh month.

* For various refinements see Kuznets, "Seasonal Variations in Industry and Trade." Also see p. 218, below.

3. Divide the data (Y) by the moving average (M_m), month by month, in order to obtain the seasonal percentages ($S\%$).

4. Tabulate the seasonal percentages for each month and find the appropriate monthly mean. Usually an average of three or four median items will be found most suitable. The twelve averages thus found constitute a crude seasonal index.

5. Center the crude seasonal index thus obtained by dividing each item in the index by the average of all the items, in order to make the annual level the base, or 100%. The results expressed as percentages constitute the required index of seasonal variations.

The analysis illustrated.—The process of time series analysis as applied to quarterly data may be briefly illustrated by the use of data representing the prevailing commercial interest rates in the United States per quarter during the years 1909 to 1913. The seasonal index is computed in Example 54. As outlined above, the process requires, first, the calculation of a centered moving average; the data are then divided by these moving averages to obtain the seasonal percentages. The seasonal percentages are next averaged by quarters. In this averaging it is usual to take the median, or preferably the mean of a few median items, the number of items included in this mean depending upon the extent of the data. This procedure eliminates the more variable percentages at each extreme, and bases the average upon several of the more stable central percentages. The resulting crude index will show the relative seasonal change, but it is desirable to express the same ratio of change in such a way that the base of the index is 100. This is readily done by dividing each item in the crude index by the average of all the items in this index.* The result is an index of seasonal variations.

Example 54.—Index of seasonal variations of commercial interest rates by quarters, 1909 to 1913. A centered moving average (M_m) is first computed (e.g., 1st $M_m = [3.8 + 2(3.9) + 2(4.2) + 2(5.5) + 4.7] \div 8 = 4.4625$; 2nd $M_m = [3.9 + 2(4.2) + 2(5.5) + 2(4.7) + 4.8] \div 8 = 4.6875$, etc.). The seasonal percentages ($S\% = Y/M_m$), as then found, are averaged by quarters to obtain the crude index. In this case the median is used, but with ampler data the average of a few median items is taken. The crude index is next divided by its own average to make the base 100%.

* Sometimes the seasonal index is centered so that the geometric mean rather than the common average is unity. This is done by finding the geometric mean and reducing the index to this mean as a base (100%). The purpose in so doing is to stabilize the normal as compared with the trend. The trend later is multiplied by the seasonal index, and if the seasonal index has a geometric mean of unity the trend and normal will have the same geometric mean.

Commercial interest rates (Y) by quarters, United States, 1909–1913

Time	1909	1910	1911	1912	1913
January–March	3.8	4.7	4.1	3.9	5.3
April–June	3.9	4.8	3.6	4.2	5.7
July–September	4.2	5.6	4.2	5.1	6.0
October–December	5.5	5.3	4.3	6.0	5.8

Centered annual moving average (M_m) of interest rates

Time	1909	1910	1911	1912	1913
January–March	....	4.98	4.48	4.26	5.64
April–June	....	5.12	4.18	4.59	5.72
July–September	4.46	5.02	4.02	4 98	
October–December	4.69	4.80	4.08	5.34	

Seasonal percentages ($S\% = Y/M_m$), and index

Time	1909	1910	1911	1912	1913	Crude index (Md)	Index
Jan.–March		94.4	91.5	91.5	94.0	92.75	92.7
April–June		93.8	86.1	91.5	99.7	92 65	92.6
July–Sept.	94.2	111.6	104.5	102.4	....	103.45	103.4
Oct.–Dec.	117.3	110.4	105.4	112.4	....	111.40	111.3
						4)400.25	4)400.0
						AM = 100.0625	100.0

Tabulating the seasonal percentages.—The data here employed are too incomplete for tabulation, but the form of such tabulation may be indicated as in Table 12. Each item is entered in its class by abbreviating the year (the last figure of the date is here used), in order to reveal progressive changes in the seasonal, if present, and the degree of concentration.

TABLE 12

Distribution of seasonal percentages by quarters

$S\%$	First	Second	Third	Fourth
115–120	....	....	....	9
110–115	....	....	0	0; 2
105–110	....	....	....	1
100–105	....	....	1; 2	
95–100	....	3		
90– 95	0; 1; 2; 3	0; 2	9	
85– 90	....	1		

As far as the data go, they do not show any marked change in the seasonal from year to year; but they do show a close concentration in the first quarter, and a wide and unsatisfactory scatter in the third.

If it is discovered that the seasonal variability is progressively changing, as for example if its amplitude is gradually lessening, then an average seasonal such as has been described is of little value. In such cases the seasonal index used currently may best be based on a very few preceding years, or some method of estimating the degree of change at any given time may be adopted.

The cycle in annual data.—It has already been shown that in annual data cyclic changes (including accidental results) may be isolated by removing the secular trend. The trend may be removed either by subtraction ($d = Y - T$) or by percentages (percentage cycle: $C\% = Y/T\%$; and $d = C\% - 100\%$). For purposes of comparison, the cycle expressed as deviations (d) may then be reduced to units of the average deviation or the standard deviation (cf. Example 55).

Example 55.—A method of measuring the cycle in annual data briefly illustrated by the data of Example 38, p. 136, where a straight-line trend was fitted. The trend is assumed to be the statistical normal; the data are compared with it ($C\% = Y/T$); deviations are obtained from 100% or from $\Sigma(C\%)/n$; and these deviations are divided by their own average deviation. The result is the average deviation cycle. The standard deviation cycle might similarly be obtained by dividing the deviations by their own standard deviation.

Year	Y	$T = N$	$C\% = Y/N$	$d\% = C\% - 100\%$	d/AD
1901	80	84	95.24	−4.76	−0 998
1902	90	86	104.65	4.65	0.975
1903	92	88	104.55	4.55	0.954
1904	83	90	92.22	−7.78	−1.631
1905	94	92	102.17	2.17	0.455
1906	99	94	105.32	5.32	1.115
1907	92	96	95.83	−4.17	−0.874
			699.98	−0.02	−0.004
				$\Sigma'd\%' = 33.40$	
				$AD = 4.7714$	

The trend and normal.—The straight-line trend is implied or used in all the examples of this chapter. As a matter of fact, this type of trend is most commonly employed in time series analysis because, as a general rule, the period of time covered is too short to indicate clearly any other trend. In some cases, however, a modified geometric trend, or some other variety, is found more satisfactory, and may be substituted in the analysis. Abbreviated approximations like that first described, however, are adapted to the straight-line trend only.

The fitting of a trend in time series analysis is often a difficult matter because of the irregularities of the cycle. The data are plotted over as long a period as possible, and a period may then be chosen as the base

of the analysis in such a way that the trend will approximate as closely as possible the long-time secular trend.

With seasonal data the normal is the trend as modified by the average seasonal movement. Sometimes, however, the seasonal is removed from the data directly, ordinarily by dividing by the seasonal index, in which case the trend may be regarded as the normal. But usually the normal is taken as the product of the seasonal index and the trend. In this form it may be used as a basis of comparison with the data (cf. Example 56).

Example 56.—Computation of the statistical normal for data of Example 54. (I) The trend is first found from the annual averages as $T = a + bx = 4.80 + 0.24x$ (point of origin July 1, 1911). (II) This trend is adjusted to the quarterly data. The trend for the first quarter of the first year (1909) is $T = a - b(ns - 1) \div (2s) = 4.80 - 0.24 \times (5 \times 4 - 1)/(2 \times 4) = 4.23$, with an increase per quarter of $b/s = 0.24/4 = 0.06$, where s means the number of divisions in the year (that is, $s = 4$ for quarterly data, or 12 for monthly data). By successive additions ($4.23 + 0.06 = 4.29$; $4.29 + 0.06 = 4.35$; etc.) the trend is extended over the 5 years. (III) The trend is multiplied by the seasonal index (see Example 54). First quar. = 92.7; 2nd quar. = 92.6; 3rd quar. = 103.4; and 4th quar. = 111.3%, each trend item being multiplied by its appropriate seasonal.

Data(Y)

Quarter	1909	1910	1911	1912	1913
January–March	3.8	4.7	4.1	3.9	5.3
April–June	3.9	4.8	3.6	4.2	5.7
July–September	4.2	5.6	4.2	5.1	6.0
October–December	5.5	5.3	4.3	6.0	5.8
$AM =$	4.35	5.10	4.05	4.80	5.70

I. Computation of trend from annual data.

Year	Y	x	x^2	xY
1909	4.35	−2	4	−8.7
1910	5.10	−1	1	−5.1
1911	4.05	0	0	0
1912	4.80	1	1	4.8
1913	5.70	2	4	11.4
	5)24.00		10	) 2.4
	$a = 4.80$			$b = 0.24$

$T = a + bx = 4.8 + 0.24x.$

II. Trend (T) of interest rates, by quarters, 1909–1913.

Quarter	1909	1910	1911	1912	1913
January–March	4.23	4.47	4.71	4.95	5.19
April–June	4.29	4.53	4 77	5.01	5.25
July–September	4.35	4.59	4.83	5.07	5.31
October–December	4.41	4.65	4.89	5.13	5.37

III. Computed normal ($N = TS$) of interest rates, 1909–1913, where the seasonal index (S) by quarters = 92.7, 92.6, 103.4, 111.3% (see Example 54).

Quarter	1909	1910	1911	1912	1913
January–March	3.92	4.14	4.37	4.59	4 81
April–June	3.97	4.19	4.42	4 64	4 86
July–September	4.50	4.75	4.99	5.24	5.49
October–December	4.91	5.18	5.44	5.71	5.98

The cycle in seasonal data.—After the normal has been obtained, the cycle may be calculated by comparing the data with the normal (usually, Y/N). This calculation has the effect of eliminating both seasonal and growth factors and leaving only a measure of the so-called cycle, including such accidental variations as may have influenced the data during the period in question. Sometimes it is possible to eliminate more or less arbitrarily the influence of the more extreme accidental occurrences, such as large-scale strikes, but in general the cycle will necessarily include much that is accidental; in fact, according to some theorists, it expresses little more than an aggregate of errors and accidents in the attempts of business to achieve coordination.

The ratio of the data to the normal gives a percentage cycle, but for purposes of comparison it is desirable to change the scale to one which will make the cycle comparable with that of another series. This may be done by first expressing the cycle in terms of deviations from 100%. It will sometimes be found, however, that the percentage cycle fails to center at 100%; that is, $\Sigma(C\%)/n$ fails to equal 100. Ordinarily the discrepancy, if not the result of numerical mistakes, may be disregarded; but if it is significant, the deviations may be taken from $\Sigma(C\%)/n$ rather than from 100%, thus making the deviations balance. If the percentage deviation cycle thus found is divided by its own average deviation, it will have an average deviation of unity, and will then be comparable to other cycles similarly computed. Such cycles may then be compared by graphic or mathematical methods. The standard deviation may be used as an alternative to the average deviation in computing the cycle, but it is not so convenient to calculate, and is subject to the criticism that it emphasizes too heavily the extreme deviations. The process of computing the cycle, including the finding of both average deviation and standard deviation cycles, is indicated in Example 57.

Example 57.—Calculation of the cycle as a percentage of the data compared with the statistical normal (cf. Example 56, p. 198). The data and normal are first given and (I) their ratio in percentages is taken. (II) Deviations are found from the average of the percentages (usually taken as 100%) in order to express the cycles as percentage deviations from normal, and the absolute average of these deviations is found. (III) The percentage cycle is then divided by the average

deviation to obtain the average deviation cycle, a form suitable for comparison or correlation with other cycles. (IV) An alternative to the average deviation cycle is the standard deviation cycle, obtained by dividing $d\%$ by $\sigma = (\Sigma d^2/n)^{1/2}$.

Data, interest rates, Y (cf. Example 54, p. 195)

Quarter	1909	1910	1911	1912	1913
January–March........	3.8	4.7	4.1	3.9	5.3
April–June............	3.9	4.8	3.6	4.2	5.7
July–September........	4.2	5.6	4.2	5.1	6.0
October–December.....	5.5	5.3	4.3	6.0	5.8

Normal, $N = ST$ (cf. Example 56, p. 198)

Quarter	1909	1910	1911	1912	1913
January–March........	3.92	4.14	4.37	4 59	4.81
April–June............	3.97	4.19	4.42	4.64	4.86
July–September........	4.50	4 75	4.99	5.24	5.49
October–December.....	4.91	5.18	5.44	5.71	5.98

I. Percentage cycle, $C\% = Y/N$. ($\Sigma C\%/n = 99.95$.)

Quarter	1909	1910	1911	1912	1913
January–March........	96.94	113.53	93.82	84.97	110.19
April–June............	98.24	114.56	81.45	90 52	117.28
July–September........	93.33	117.89	84.17	97.33	109 29
October–December.....	112.02	102.32	79.04	105.08	96.99

II. Percentage deviation cycle, $d\% = C\% - 99.95\%$.

Quarter	1909	1910	1911	1912	1913
January–March........	− 3.01	13.58	− 6.13	−14.98	10.24
April–June............	− 1.71	14.61	−18.50	− 9.43	17.33
July–September........	− 6.62	17 94	−15.78	− 2.62	9.34
October–December.....	12 07	2.37	−20.91	5.13	− 2.96

III. Average deviation cycle, d/AD. ($AD = 10.2630$.)

Quarter	1909	1910	1911	1912	1913
January–March........	−0.293	1.323	−0.597	−1.460	0.998
April–June............	−0.167	1.424	−1.803	−0.919	1.689
July–September........	−0.645	1.748	−1.538	−0.255	0.910
October–December.....	1.176	0.231	−2.037	0.500	−0.288

IV. Standard deviation cycle, d/σ. ($\sigma = 11.9176$.)

Quarter	1909	1910	1911	1912	1913
January–March........	−0.253	1.139	−0.514	−1.257	0.859
April–June............	−0.143	1.226	−1.552	−0.791	1.454
July–September........	−0.555	1.505	−1.324	−0.220	0.784
October–December.....	1.013	0.199	−1.755	0.430	−0.248

A summary of time series analysis.—In order to give a more unified view of the process of time series analysis as previously described,

the three preceding problems are brought together in a somewhat different form in Example 58. This form will probably be found most convenient for actual use. For example, suppose that an analysis is to be made of monthly data covering ten years. Large sheets of paper (for example, 17 in. by 22 in.), with plain ruling, 6 squares to the inch, may be used. Data are preferably written without spaces between years in order to facilitate the finding of the moving average.* The moving average is computed, and the seasonal percentages are obtained as a ratio of the data to the moving average (Y/M_m) and are usually sufficiently accurate if written as a percentage with one decimal. The seasonal index may then be separately calculated by listing these percentages by months, ranking them, averaging the central items, and centering as before. The seasonal index thus obtained is written in the S column, repeated for each year. The trend is then calculated as previously explained and multiplied by the seasonal indexes to obtain the normal. In taking the ratio of the data to the normal (Y/N), it is usually advisable to carry the percentages to two decimal places; otherwise, deviations from 100% (or from $\Sigma C\%/n$) may have too few significant places. The deviation cycle thus obtained is then reduced to units of its own average deviation.

Example 58.—Summary of time series analysis of interest rates (Y) by quarters, United States, 1909–1913 (cf. Examples 54, 56, and 57). The successive columns are as follows: $M\Sigma$ = annual moving total of Y, 5 quarters doubling 3 central quarters; $M_m = M\Sigma/8$, annual moving average centered; $S\% = Y/M_m$; S = seasonal index (cf. Example 54); T = straight-line trend computed from annual averages (cf. Example 56); $N = ST$; $C\% = Y/N$; $d\% = C\% - \Sigma(C\%)/n$, or usually $C\% - 100\%$; $d/AD = d\% \div \Sigma'd\%'/n$. The footings of the columns provide certain obvious checks.

* The moving totals, if they are to be centered, may be computed by placing on the adding machine the first January Y, plus twice the February Y, plus twice the March Y, etc., . . . plus twice the December Y, plus the January Y, and taking a sub-total. The first January and February Y's are then subtracted once each, the January and February Y's of the next year added once each, and another sub-total taken. Then similarly the first February and March Y's are subtracted once each, and the second year February and March Y's are added once each, and a sub-total is again taken. This procedure may be carried out to the end of the series. The first moving total centers in July. The moving totals divided by 24 will give the annual centered moving averages. As a rule, however, it is sufficiently accurate to take a straight 12 months' moving average centered in the seventh month. In this case the first 12 months Y's are placed on the adding machine and sub-totaled; the first January is subtracted and the second January added, and another sub-total taken; and so on, to the end of the series. In taking the moving total it is well to check occasionally by reading the tape to make sure an error has not entered into the calculation; otherwise, an error near the beginning will invalidate the whole operation.

Time		Y	$M\Sigma$	M_m	$S\%$	S	T	N	$C\%$	$d\%$	d/AD
Year	Quar.										
1909	1	3.8	....	....	.....	92.7	4.23	3 92	96.94	− 3.01	−0.293
	2	3.9	....	....	.....	92 6	4 29	3.97	98 24	− 1.71	−0.167
	3	4 2	35.7	4.46	94.2	103.4	4.35	4.50	93.33	− 6 62	−0.645
	4	5.5	37.5	4.69	117.3	111 3	4.41	4.91	112 02	12.07	1.176
1910	1	4.7	39.8	4.98	94.4	92 7	4.47	4.14	113.53	13.58	1.323
	2	4.8	41.0	5.12	93.8	92.6	4.53	4.19	114.56	14.61	1.424
	3	5.6	40.2	5.02	111.6	103.4	4.59	4.75	117.89	17 94	1.748
	4	5 3	38 4	4.80	110 4	111 3	4.65	5.18	102.32	2 37	0.231
1911	1	4.1	35.8	4.48	91 5	92 7	4.71	4 37	93.82	− 6 13	−0.597
	2	3 6	33.4	4.18	86.1	92.6	4.77	4.42	81.45	−18.50	−1.803
	3	4.2	32.2	4.02	104.5	103.4	4.83	4.99	84.17	−15 78	−1.538
	4	4.3	32 6	4.08	105 4	111 3	4.89	5 44	79.04	−20.91	−2 037
1912	1	3.9	34.1	4 26	91 5	92.7	4.95	4 59	84.97	−14.98	−1.460
	2	4.2	36.7	4 59	91 5	92 6	5.01	4.64	90.52	− 9 43	−0.919
	3	5.1	39.8	4.98	102 4	103 4	5.07	5.24	97.33	− 2.62	−0.255
	4	6.0	42.7	5.34	112.4	111 3	5 13	5.71	105.08	5.13	0.500
1913	1	5.3	45.1	5.64	94.0	92.7	5.19	4 81	110.19	10.24	0.998
	2	5 7	45.8	5.72	99.7	92.6	5.25	4 86	117.28	17.33	1.689
	3	6 0	....	....	......	103.4	5.31	5.49	109 29	9.34	0.910
	4	5.8	....	....	......	111.3	5.37	5.98	96.99	− 2.96	−0.288
		96.0		16)1600 7			96.00	96.10	20)1998.96	− 0.04	−0.003
					100 04				99.95		
										102.61	9.999
										−102.65	−10.002
									$\Sigma'd\%'$ =	205.26	20.001
									AD =	10.2630	

The columns may be footed in pencil, so as not to interfere with later extensions, and certain obvious checks applied. In the first place, $\Sigma Y = \Sigma T$; also $\Sigma S\%$ should average approximately 100. The seasonal index should be adequately checked before writing it in the S column. The normal should foot approximately to ΣY or ΣT, but a discrepancy may appear owing to irregularities in the seasonal and trend.* The percentage cycle should average approximately 100, and

* The discrepancies in the footings, aside from the slight irregularities arising from the rounding of decimals, are caused chiefly by minor inconsistencies in the logic of the analysis. If the distribution of the data is strictly arithmetic, the seasonal should be computed from deviations ($Y - M_m$) rather than ratios (Y/M_m); the trend plus the seasonal would be taken as the normal; and the cycle would be the data less the normal ($Y - N$). Close checks could then be applied, although the moving average would not necessarily foot exactly the same as the data. But

as a general rule the average is rounded to 100 in taking the deviations. The footing of the $d\%$ column should check against the footing of the per cent cycle, since $\Sigma(d\%) = \Sigma(C\%) - 100n$, or $\Sigma(d\%) = n$ times the error in rounding $\Sigma(C\%)/n$. The absolute sum of the $d\%$ column divided by n gives the average deviation (AD). The footing of the

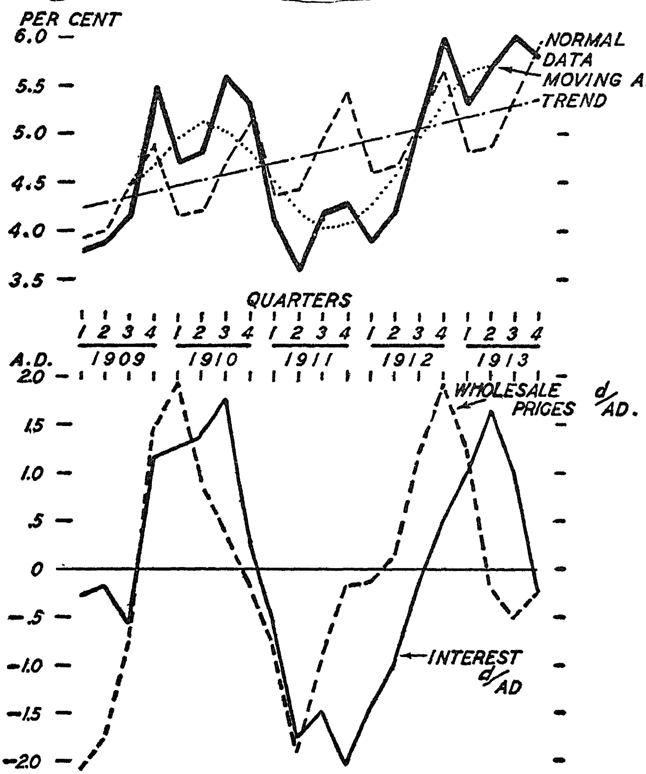


CHART 28

Time series analysis of commercial interest rates in the United States by quarters, 1909–1913, showing the data, the moving average, the trend, the normal (upper figure), and the average deviation cycle derived from this analysis, together with an analogous cycle of wholesale prices (lower figure). It will be seen that the normal is in effect an averaging of the data including its seasonal and trend but eliminating its cycle. The cycle is, therefore, obtained by comparing the data with the normal. The use of the average deviation scale (lower figure) makes the interest rates and wholesale prices mathematically comparable. For the computation see Example 58.

d/AD column, times AD, should balance the total of the $d\%$ column, and should give an absolute total of n. Some of these checks are obvi-

if the distribution of the data were strictly logarithmic, this procedure should be applied to the logs of the data and the trend. Practically, however, the usual procedure is justified.

ously subject to slight variations occasioned by the rounding of the decimals. If the standard deviation cycle is to be used, the squares of $d\%$ may be listed from Barlow's "Tables" on an adding machine, or accumulated by multiplication on a calculating machine. The squares are averaged and the square root taken as σ, which is used as a divisor in place of AD. As previously indicated, however, the average deviation is more convenient for ordinary graphic or correlation comparisons, but the standard deviation may prove to be more convenient if multiple correlations are to be developed.

Short cuts.—There are a number of short cuts which will readily be suggested by experience in time series analysis. In the first place, unless the moving average is to be plotted, the seasonal index may be computed from the ratios of the data to the moving totals ($Y/M\Sigma$), in which case the factor $\frac{1}{12}$ must be eliminated in the final adjustment. In the second place the normal may be advantageously computed, with less danger of error, by the following method. After the seasonal index has been derived, and the trend equation for n years ($Y = a + bx$) ascertained, the trend items for the first year of the series may be obtained by the equations: *

$$T_1 = a - b(ns - 1) \div (2s)$$

$$T_2 = a - b(ns - 3) \div (2s)$$

$$T_3 = a - b(ns - 5) \div (2s)$$

$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$

$$Ts = a - b(ns - 2s + 1) \div (2s)$$

where s is the number of subdivisions to the year ($s = 4$ for quarterly data and $s = 12$ for annual data). The trend items for the first year in the series thus obtained are multiplied by the seasonal index to obtain the normal for the first year in the series. The normal for succeeding years may then be written by adding successively to the first item b times the seasonal for that month (or quarter); to the second item b times the seasonal for that month (or quarter), etc. These successive additions may readily be made on the calculating machine with little chance of error. The results will be exactly the same as those previously obtained. The process appears complicated when first described, but in actual practice will really be found easy and rapid (cf. Example 59).

Example 59.—Alternative method of calculating the normal from quarterly data of Example 54. After the seasonal index (1st quar. 92.7; 2nd quar. 92.6;

* This method was developed by Mr. Floyd B. Haworth of the University of Iowa.

3rd quar. 103.4; 4th quar. 111.3) and the trend equation based on the average data by years ($T = 4.80 + 0.24x$) have been obtained (cf. Examples 54 and 56), the trend for the first quarter in the first year, in the series (1909), is calculated by the equation $T = a - b(ns - 1) \div (2s)$ where s is the number of subdivisions to the year, a and b are the constants of the trend equation, and n is the number of years. The trend items for succeeding quarters of 1909 are computed as indicated. The normals for succeeding years are obtained by adding successively to $N = TS$, 1909, the quantities bS, quarter by quarter, as indicated.

1909 Quarter		T	$\times\ S$	$=\ N$
Jan.–Mar.	$T = a - b(ns - 1) \div (2s) = 4.80 - 0.24\left(\frac{19}{8}\right)$	$= 4.23$	92.7	3.92121
Apr.–June	$T = a - b(ns - 3) \div (2s) = 4.80 - 0.24\left(\frac{17}{8}\right)$	$= 4.29$	92.6	3.97254
July–Sept.	$T = a - b(ns - 5) \div (2s) = 4.80 - 0.24\left(\frac{15}{8}\right)$	$= 4.35$	103.4	4.49790
Oct.–Dec.	$T = a - b(ns - 7) \div (2s) = 4.80 - 0.24\left(\frac{13}{8}\right)$	$= 4.41$	111.3	4.90833

(The increase in T is obviously $b/4$, which may be added successively to the first item in the T column, 4.23, to obtain the other items in that column.)

1909 N	bS	1909 $N + 0bS$	1910 $N + 1bS$	1911 $N + 2bS$	1912 $N + 3bS$	1913 $N + 4bS$
3.92121	0.22248	3.92	4.14	4.37	4.59	4.81
3.97254	0.22224	3.97	4.19	4.42	4.64	4.86
4.49790	0 24816	4.50	4.75	4.99	5.24	5.49
4.90833	0.26712	4.91	5.18	5.44	5.71	5.98

Projecting the normal.—After a time series has been analyzed for a given base period, say for eight or ten years, current items may be calculated and evaluated with reference to the cycle by projecting the normal on the basis of the trend and seasonal previously found. Thus in Example 58, p. 201, the trend may be extended by adding $b/4$ successively, giving for the successive quarters of 1914 the items 5.43, 5.49, 5.55, and 5.61. These trend items may be multiplied by their respective seasonals, to obtain the following successive normals: 5.03, 5.08, 5.74, and 6.24. The data, as they are reported month by month or quarter by quarter, may be divided by the normal thus projected to obtain the percentage cycle, which may be reduced to the average deviation cycle, as before, without recomputing the average deviation. It is understood, of course, that the " data " mean the figures as reported, adjusted as the preceding data were adjusted, by deflation or reduction to daily averages, or otherwise. By this procedure the study may be projected for current data, and an estimate of business conditions with respect to the cycle may be obtained.

It is obvious that if the normal is projected very far from the base period it will become less and less reliable, inasmuch as the seasonal, trend, and average deviation may gradually change. Hence it is a common practice to project the normal about a year, and then rework the whole time series analysis through a new base period. This new

period may be the old one enlarged, or otherwise modified so as to secure a reasonable trend. Sometimes, however, it will be found satisfactory to project the normal in such a way as to modify the seasonal and trend from time to time without reworking the whole base period. If the tabulation is kept up to date by entering the new items, extending the moving average, and computing the seasonal percentages, the seasonal index may be quickly revised at any time. It is even possible to compute a new seasonal index number for the current month by merely averaging in the usual way the seasonal percentages for the last several like months, though this would eliminate the possibility of centering the seasonal index. The trend likewise may be readily recomputed for a revised period by means of the moving averages, which when taken a year apart are annual averages. By reference to a chart of the data it may be judged what period would be practicable to use as the base of the readjusted trend. If no marked cyclic change is taking place it might be practicable to base each new trend item on the last ten years (or other base period) ending at the current date. This may be done by selecting n moving averages a year apart, ending with the last moving average available. If these moving averages had been centered (e.g., thirteen items in monthly data; two end items weighted one-half), the trend item required would be

$$T = a + b(n/2)$$

If the method of semi-averages is used, the trend item would then simply be, if n were an even number,

$$T = (3S_2 - S_1)/n$$

If, however, n is an odd number, the formula becomes

$$T = (S_2 + S_1 + M)/n + (S_2 - S_1)[2n/(n^2 - 1)]$$

where M is the middle moving average, between S_1 and S_2, and the other symbols are as used in the method of semi-averages.

These formulas may be illustrated with reference to the data of interest rates, Example 58. Suppose that in 1914 the interest rate for the first quarter, 4.2%, had just become available, and it was desired to evaluate this figure on the basis of four years ending with that quarter. The centered moving average may now be extended, resulting in the figure 5.56, which centers in the third quarter of 1913. The four moving averages in recent third quarters, ending with the moving average just mentioned, are 5.02, 4.02, 4.98, and 5.56. The formula may now be solved, as follows:

$$T = (3S_2 - S_1)/n = [3(5.56 + 4.98) - (4.02 + 5.02)]/n = 5.645$$

The computation may quickly be listed on the adding machine as follows:

-5.02	M_m 3rd quarter	1910
-4.02	M_m 3rd quarter	1911
4.98	M_m 3rd quarter	1912
4.98		
4.98		
5.56	M_m 3rd quarter	1913
5.56		
5.56		
$4)\overline{22.58}$	$= 3S_2 - S_1$	
5.645	$= T$, 1st quarter	1914

This is the trend item for the first quarter of 1914 on a four-year basis (four years ending the middle of the final quarter), and this multiplied by the seasonal (92.7%) gives a normal of 5.233. The percentage cycle (80.26) and the average deviation cycle (-1.92) may now be obtained as before.

If, however, the new item for the first quarter of 1914 were to be evaluated on the basis of five years ending the middle of that quarter, the trend would be found by the second formula given above. In that case the moving averages involved would be: 4.46, 5.02, 4.02, 4.98, 5.56, and the formula would be solved as follows:

$$\begin{aligned} T &= (S_2 + S_1 + M)/n + (S_2 - S_1)[2n/(n^2 - 1)] \\ &= 24.04/5 + 1.06 \times 0.4167 = 5.250 \end{aligned}$$

The cycle could now be measured as before.

The moving average of non-centered data.—If in monthly data a twelve-term moving average is taken, formulas such as have just been stated will yield trend items centered at the end of the final month instead of the middle of that month. Hence they must be corrected by subtracting $b/24$ (or in general $b/2s$). They now take the form:

$$T = a + b(ns - 1)/2s$$

which becomes, when n is an odd number:

$$T = (3S_2 + M - S_1) \div n + [(s - n)/n][2(S_2 - S_1) \div s(n^2 - 1)]$$

and when n is an even number:

$$T = (3S_2 - S_1) \div n - 2(S_2 - S_1) \div sn^2$$

In applying these equations it is perhaps about as easy to solve the trend equation $T = a + bx$ and use the form first given. However,

the second equation (n an odd number) will prove very useful if the second term $[(s - n)/n][2(S_2 - S_1)/s(n^2 - 1)]$ may be neglected, as is usually the case with monthly data. When the last equation (n an even number) is applied to monthly data, the last term, $2(S_2 - S_1) \div s(n^2 - 1)$ expresses a half month's trend, which is usually negligible unless the trend is very marked. Hence, these equations may be found useful as approximations for projecting the trend or for frequent revisions.

A moving base.—As has been indicated, it is a very difficult problem to secure a base period that is typical with respect to the trend. For this reason there is something to be said in favor of a base period consisting of a fixed number of years prior to the current date, particularly when working with the rather irregular fluctuating data available in smaller areas, such as a city or a state. It is quite easy in such an analysis to set up a moving base, of perhaps eight or ten years, and to revise the trend month by month and the seasonal semi-annually. If such an analysis is always stated to be with reference to the given base period, and if the data are so charted as to exhibit the nature of the base period, there need be no misunderstanding as to the results obtained. However, severe cyclic changes would invalidate this procedure. In the case of a severe depression the trend could be tentatively stabilized as soon as it had reached a horizontal position, and so continued until a reworking of the data on the basis of a longer secular trend could be undertaken. In any case, the relation of the short working trend to the longer secular trend should always be kept in mind. It is obvious that a study of long-time trends, and an adjustment of the trend actually used with reference to such a study, will doubtless give the most dependable results.

The composite cycle.—If several related statistical series have been analyzed and the cycles computed, they may be combined by a process of averaging in order to give a picture of the general cyclic change in the whole field which they are assumed to sample. If the series taken together represented the whole field of final production, they might be combined by a simple arithmetic mean of the percentage cycle or the percentage deviations. But, as a rule, they represent merely samples which, in a measure, overlap with each other, as in the case of crops and car loadings, or iron and automobile production. Hence, it is customary to combine the average or standard deviation cycles rather than the percentage cycles, thus eliminating the variability with respect to the amplitudes. In computing the average, weights are commonly used representing an estimate of the importance of each series as a measure of the cycle. Such weights are difficult to determine since they take

into consideration the quantity or value of the series, and also the validity or accuracy of the figures in the series. But after these weights have been determined, their application to the measure of the cycle in question will give a weighted average representing the composite cycle. The cycle thus obtained is preferably left in average deviation units, since, strictly speaking, it cannot be made to yield a very close estimate of the real percentage change. However, it may be reduced to a percentage estimate by multiplying by the harmonic mean of the average deviations, weighted by the estimated importance of the series. The harmonic mean is used because the average deviations enter inversely in the process of combining. The procedure is illustrated in Example 60. It will be seen that the whole process may be abbreviated as indicated in the second part.

Example 60.—Calculation of the composite cycle in average deviation units for five series listed in the *Annalist Index of Business Activity* (January, 1931). The percentage cycle was computed by the *Annalist* by methods similar to those previously described. The deviations are taken from 100%, and are divided by the average deviation of each series as reported in previous studies of the *Annalist*. The resulting ratios, d/AD, are averaged by the use of weights representing the *Annalist's* estimate of the relative importance of each series. The resulting composite, $-3.24AD$, represents the degree to which business was below normal in units of average deviation on the given date.

	% cycle	d%	AD	d/AD	Wt.	Product
Pig iron...........	55.0	−45.0	20%	−2.25	10	−22.5
Freight car loadings	79.1	−20.9	5%	−4.18	20	−83.6
Electric power prod.	83.8	−16.2	4%	−4.05	10	−40.5
Automobile prod....	62.6	−37.4	22%	−1.70	10	−17.0
Cotton consumption	71.7	−28.3	9%	−3.14	15	−47.1
		HM_w =	7.13%	5)15 32	65	)−210 7
				3.06	Comp. cyc. =	− 3.24AD
					Comp. AD =	7.13%
					Comp. cyc. =	− 23.1%

The composite cycle in percentage units may be obtained more quickly as follows:

	% cycle	Wt./AD	Adj. wt.	Product
Pig iron.........................	55.0	$\frac{10}{20}$ =	0.50	27.5
Freight car loadings..............	79.1	$\frac{20}{5}$ =	4.00	316.4
Electric power production.........	83.8	$\frac{10}{4}$ =	2.50	209.5
Automobile production............	62.6	$\frac{10}{22}$ =	0.45	28.17
Cotton consumption..............	71.7	$\frac{15}{9}$ =	1.67	119.74
	HM_w of AD's = 65 ÷ 9.12		9.12	)701 31
	= 7.13AD		Comp. cyc. =	76.9%
				or − 23.1%

$$-23.1\% \div 7.13AD = -3.24AD$$

The calculation of the composite cycle is more fully illustrated in the data of Example 61, where the same five series previously used have been combined into a tentative composite cycle for the years 1929 and 1930, using the *Annalist's* average deviations and weights.

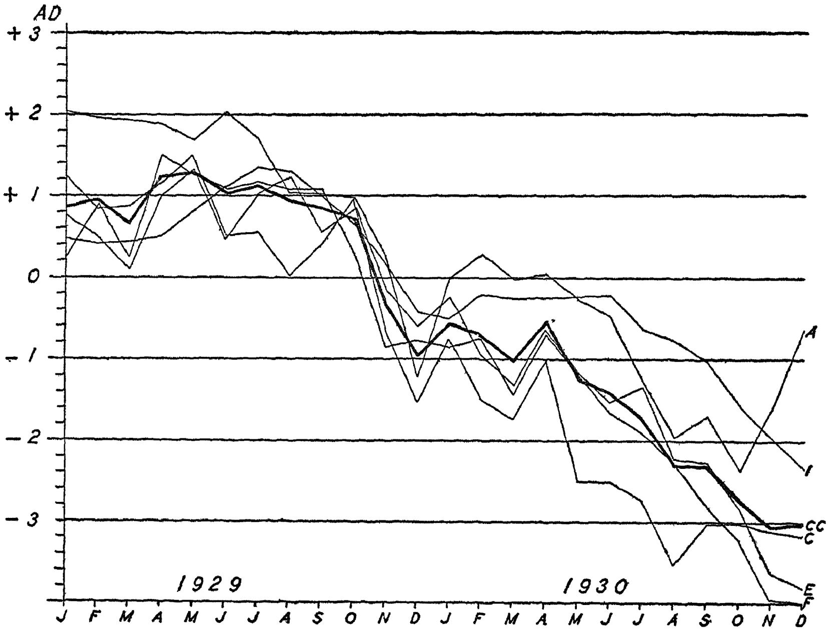


CHART 29

The composite cycle derived tentatively from five individual percentage cycles as published by the *Annalist*, monthly, 1929–1930; the series being, pig iron (*I*); freight car loadings (*F*); electric power production (*E*); automobile production (*A*); and cotton consumption (*C*). The composite cycle (*CC*) is represented by a heavy line. The method of making the composite is illustrated in Example 61.

The result is pictured in Chart 29, which shows the individual series and the composite derived from them.

Example 61.—Illustration of computation of composite cycle (*CC*) in units of average deviation (*AD*), based upon five individual percentage cycles (*C*%) published by the *Annalist*, monthly, 1929–1930 (cf. Chart 29). The individual cycles were obtained by comparing the data with the statistical normal by methods similar to those described in Example 57. The calculation of the pig iron deviations (*d*%) and average deviation cycle (*d*/*AD*) for 1929 and the composite of pig iron and four analogous series for January, 1929, is here given: the remainder of the calculation is left as a laboratory exercise for the student. The weights employed represent the *Annalist's* estimate of the relative importance of the series as related to the measurement of the business cycle in the United States.

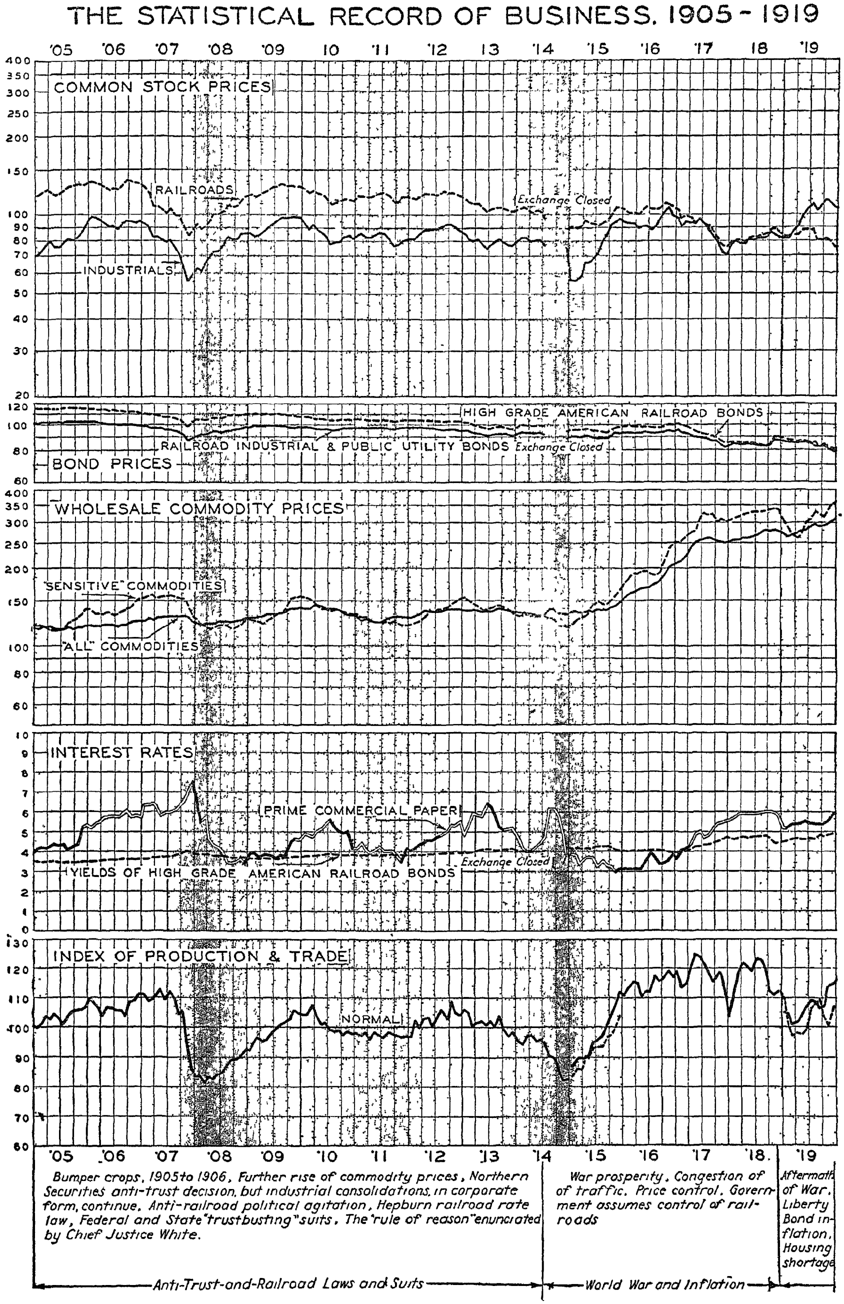


CHART 29a

Cycles of production and trade, and other indexes of the business cycle, 1905–1919, illustrating a novel and effective method of distinguishing the cycles by shading. Reprinted by permission from Persons, W. M., "Forecasting Business Cycles," John Wiley & Sons, New York, 1931, p. 129.

Pig iron

1929	C%	d%	d/AD
January	109.6	+ 9.6	+0.480
February	108.7	+ 8.7	+0 435
March	108.4	+ 8.4	+0.420
April	110.4	+10.4	+0.520
May	116.3	+16 3	+0.815
June	123.1	+23.1	+1.155
July	127.4	+27.4	+1.370
August	126 3	+26.3	+1 315
September	119.7	+19.7	+0.985
October	112.9	+12.9	+0.645
November	103.7	+ 3.7	+0.185
December	91.7	− 8.3	−0.415

Average deviation = 20
Weight = 10

Composite, January, 1929

	d/AD	Weights	Product
Pig iron	+0.480	10	4.80
Freight car loadings	+0.24	20	4.80
Electric power production	+0.775	10	7.75
Automobile production	+2 082	10	20.82
Cotton consumption	+1.244	15	18.66
		65	)56.83

$CC = 0.874AD$

Components of the cycle.—Various series entering into the business cycle show different characteristics as illustrated by Chart 30. During the years 1903–1912, as illustrated by this chart, speculative series, such as shares on the New York Stock Exchange, tended to precede the central portions of the business cycle, whereas series representing financial strain, such as interest rates, tended to lag. Since the war, however, there has been much irregularity, and no very definite precedence or lag can be made out. In addition to economic series, several other series of a more or less social nature responded to the business cycle, such as employment, crime, marriage and divorce, and suicide rates.

Forecasting the business cycle.—The problem of forecasting the business cycle is not strictly a part of the present discussion, which is concerned with methods of analyzing the data. It should be pointed out, however, that the business cycle is not periodic; the length in the United States has varied from one to nine years (cf. Chart 31), and the degree of intensity has varied fully as widely. Various mechanical methods of forecasting the cycle which have temporarily shown some

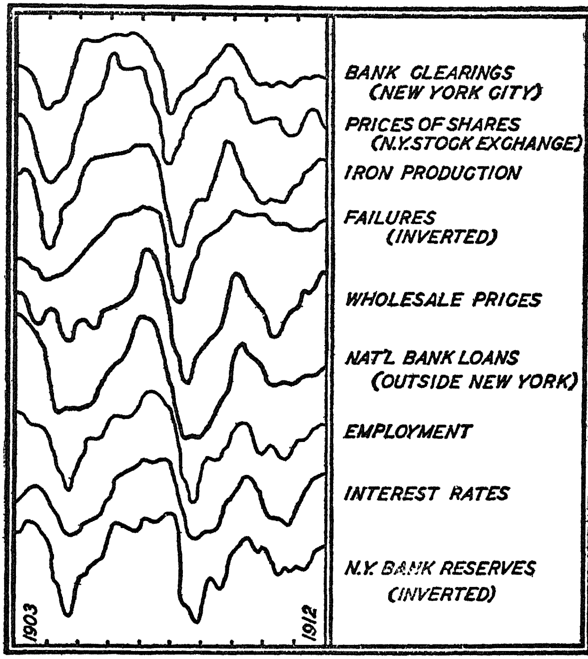


CHART 30

Elements of the business cycle, 1903–1912, developed from data compiled by the Harvard Committee on Economic Research, showing the tendency of some series to lag and others to precede the central portions of the cycle.

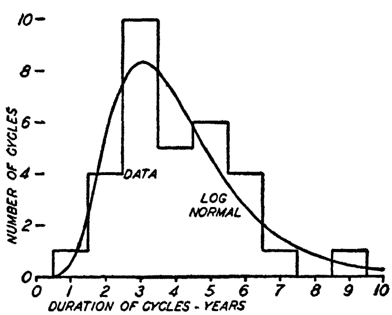


CHART 31

Business cycles in the United States, 1796–1923, classified according to length, from a study by the National Bureau of Economic Research, together with the logarithmic normal probability curve fitted to the data. The cycles vary from one to nine years with frequencies respectively as follows: 1, 4, 10, 5, 6, 4, 1, 0, 1; the normal frequencies being: 0.68; 5.21; 8.15; 6.96; 4.65; 2.77; 1.58; 0.88; 0.49. The arithmetic mean is 4.031 years, the geometric mean is 3.676 years, and the mode 3.046. (From the *Journal of the American Statistical Association*, December, 1929, p. 366, by permission.)

success have, for the most part, broken down. It is obvious that the business cycle is the product of more complicated factors than can be reduced to statistics, and that even though statistical studies may perform the invaluable task of registering the progress of events, they cannot hope to do more than furnish the basis of intelligent estimates regarding the immediate future. Within certain limits, presumably, history repeats itself, and the future may be judged from the past; but since new factors enter in, exact forecasting becomes impossible.

SUPPLEMENTARY METHODS

Logarithmic method of time series analysis.—From the theoretical point of view the most consistent method of time series analysis is one which uses the logarithms of the edited and adjusted data as the basis (Y) of the analysis. When this is done the measurement of seasonal variations and deviations from normal may be taken as differences rather than as ratios, since log differences are representative of ratios. The process may most conveniently be carried out by first eliminating the trend, and then computing and eliminating the seasonal. The procedure is indicated in Example 62, and may be briefly summarized as follows:

After the data have been suitably edited and adjusted, the logarithms are taken as Y. A trend is then fitted to these logarithms. Over a comparatively short period of years the trend adjusted to the logarithms probably will not differ appreciably from that fitted to the data, but care must be taken in projecting the trend to see that it remains consistent. Over a considerable period of years the trend fitted to the logarithms will obviously be of a different type from that fitted to the data, but if frequent readjustments are made the straight-line trend will commonly be found satisfactory; otherwise, a modified geometric or other type may be indicated.

When the trend has been calculated, it is removed from the logarithmic data (Y) by subtraction, and a moving average of the residuals is calculated. By subtracting the moving average from the detrended Y, seasonal differences are obtained measuring the seasonal variations. These may be combined into a seasonal index by the usual method; that is, the differences for like seasons are averaged, as has already been explained in previous forms of analysis (cf. Example 54, p. 195). The crude seasonal index thus obtained may be centered arithmetically; that is, the average of all the seasonal items in the crude index is subtracted from each item, to make the algebraic sum zero. The seasonal index thus determined may then be subtracted from the detrended Y,

and a series of cyclic deviations results. If an average deviation cycle is desired, it may be made directly from these logarithms by reducing them to units of their own average deviation. But if the usual percentage cycle is required, it may be found as the antilogs expressed in percentages. The cycle thus expressed in percentages is based upon the geometric mean as 100% and therefore averages a little more than unity. It may, however, be centered about the arithmetic mean by the usual procedure of dividing each item by the average of all the items, if this seems desirable. Since an s-term moving average of log $C\%$ is the same as the moving average of $Y-T$, the latter may be regarded as the smoothed cycle.

Although the method of logarithmic time series analysis is more consistent than that usually employed and is probably not any more difficult to compute, it is not in general use. It is, however, worthy of more consideration than it usually receives. When applied directly to the data rather than to the logarithms of the data, the method constitutes a fairly rapid way of obtaining a good approximate result (Cd instead of log $C\%$).

Approximating the normal for the base years.—In connection with the method of determining a projected normal on the basis of the seasonal averages and the annual averages (Example 53, p. 191), it was noted that the method may be extended to cover the entire base period. This method is illustrated in Example 63 as applied to commercial interest rates by quarters, United States, 1910–1913. Although the period is too short for a valid analysis, it is sufficient to illustrate the method. It will be seen that the data are averaged both by columns and rows,

Example 62.—Time series analysis by the logarithmic method applied to quarterly data (cf. Example 58, p. 201). The logarithms of the interest rates are taken as Y, the data for analysis. A straight-line trend is fitted to Y in the usual manner and is subtracted, giving the detrended data Y–T. A centered annual moving average (MA) is found and subtracted from Y to obtain the seasonal differences (Sd). A crude seasonal index by quarters is obtained by averaging the seasonal differences for the first quarter, for the second quarter, etc. (average of two or three median items). This crude index is centered by subtracting from each item the average of the four items, thus making the sum zero. An approximate index may be obtained more quickly by a similar procedure based on Y–T. The seasonal index is then subtracted from Y–T to obtain the cycle percentages in logarithms (log $C\%$). The antilogs constitute the seasonal percentage cycle ($C\%$) (negative logs are most conveniently expressed as supplements of 1, thus: $-0.014 = 0.986-1$, etc.). The final percentage cycle is automatically centered about the geometric mean; that is, the product of all items is unity. But it may be readily centered about the arithmetic mean, if desired, by dividing each item by the arithmetic mean. Checks on the work are indicated by the footings of the columns. The same method may be conveniently applied directly to the data to obtain approximate cycle differences (Cd).

Year	Quarter	Interest rate (I)	(log I) Y	(Trend) T	$Y-T$	MA	$(\overline{Y-T}-MA)$ Sd	$(\overline{Y-T}-S)$ log $C\%$	(Antilog) cycle
1909	1	3.8%	0.580	0.625	−0 045	....	....	−0.014	0.97
	2	3.9	0.591	0.630	−0.039	....	....	−0.006	0 99
	3	4 2	0.623	0.636	−0.013	0.009	−0.022	−0.028	0.94
	4	5.5	0.740	0.641	0.099	0.027	0.072	0.050	1.12
1910	1	4.7	0.672	0.646	0.026	0.048	−0 022	0.057	1.14
	2	4.8	0.681	0.651	0.030	0 057	−0.027	0.063	1.16
	3	5.6	0.748	0.657	0.091	0.042	0.049	0.076	1.19
	4	5.3	0.724	0.662	0.062	0.014	0.048	0.013	1.03
1911	1	4.1	0.613	0.667	−0.054	−0.023	−0.031	−0.023	0.95
	2	3.6	0.556	0.673	−0.117	−0.055	−0.062	−0.084	0.82
	3	4.2	0.623	0.678	−0.055	−0.074	0.019	−0.070	0 85
	4	4 3	0.633	0.683	−0.050	−0.074	0.024	−0.099	0.80
1912	1	3.9	0.591	0.688	−0.097	−0.060	−0.037	−0.066	0.86
	2	4.2	0.623	0.694	−0.071	−0.037	−0.034	−0.038	0.92
	3	5.1	0.708	0.699	0.009	−0.007	0.016	−0.006	0.99
	4	6.0	0.778	0.704	0.074	0.020	0.054	0.025	1.06
1913	1	5.3	0.724	0.710	0.014	0.041	−0.027	0 045	1 11
	2	5.7	0.756	0.715	0.041	0.042	−0.001	0.074	1.19
	3	6.0	0.778	0.720	0.058	....	....	0.043	1.10
	4	5.8	0.763	0 726	0.037	....	....	−0.012	0 97
			20)13.505	13.505	0	....	....	0	20.16
			0.67525						

Year	Trend based on annual Y averages	
1909	0.63350	−1.26700
1910	0.70625	−0.70625
1911	0.60625	....
1912	0.67500	0.67500
1913	0.75525	1.51050
	5)3.37625	10)0.21225
	a = 0.67525	b = 0.021225

Seasonal index

(a) Based on Sd, average of 2 median items

Jan.–Mar.	Apr.–June	July–Sept.	Oct.–Dec.
−0.031	−0.033	+0.015	+0.049

(b) Based on $Y-T$, average of 3 median items

Jan.–Mar.	Apr.–June	July–Sept.	Oct.–Dec.
−0.034	−0.032	+0.013	+0.053

thus yielding the seasonal averages (SA) and the annual averages (AM). The trend equation is computed, in this case, by the method of least squares, but the method of semi-averages might be appropriately substituted. The normal for the first year of the series is computed by a special equation as indicated, and the normals for succeeding years are obtained by adding the requisite annual increases for each quarter. The method is suitable for a rough approximation and may be measurably valid when the data are fairly regular, but it has the defects previously noted of using arithmetic means instead of medians and of failing to secure a suitable base for the measurement of the seasonals.

Example 63.—Estimation of the normal for the base years of Example 53, p. 191. The data are first averaged by columns and rows, giving the seasonal averages (SA) and the annual averages (AM). A trend equation is computed from the annual averages; the method of least squares is here used, but semi-averages might be

substituted. The seasonal slope (SS) is found for the first season by the equation $SS_1 = SA_1 \div (a/b + \overline{1 - s}/2s)$, and for the following seasons by substituting for $\overline{1 - s}$ successively $\overline{3 - s}$, $\overline{5 - s}$, etc., for the required number of seasons, where s is the number of subdivisions to the year. The normal is then computed for each season by the equation $N = SA + SSx$, where x is the time in years centered as in computing the trend. When projections are made, as previously explained, the trend is changed to include the last four years ending with the current item. In practice a more extended base should be employed.

Data (Y)

Quarter	1910	1911	1912	1913	SA
January–March	4.7	4.1	3.9	5.3	4.500
April–June	4 8	3.6	4.2	5.7	4.575
July–September	5.6	4.2	5.1	6.0	5.225
October–December	5.3	4.3	6.0	5.8	5.350
$AM =$	5.10	4.05	4.80	5.70	4.9125

I. Computation of trend (T) from annual averages.

Year	Annual averages Y	x	x^2	xY
1910	5 10	−1.5	2.25	−7.650
1911	4 05	−0.5	0.25	−2.025
1912	4.80	0.5	0.25	2.400
1913	5.70	1.5	2.25	8.550
	4)19 65		5.00	5)1.275
	$a =$ 4 9125			$b =$ 0.255

$T = 4.9125 + 0.255x$; origin, January 1, 1912.

II. Computation of normal for 1910 ($x = -1.5$).

$$SS_1 = SA_1 \div (a/b + \overline{1 - s}/2s) = 4.500 \div (19.2647 - 0.375) = 0.2382$$
$$SS_2 = SA_2 \div (a/b + \overline{3 - s}/2s) = 4.575 \div (19.2647 - 0.125) = 0.2390$$
$$SS_3 = SA_3 \div (a/b + \overline{5 - s}/2s) = 5.225 \div (19.2647 + 0.125) = 0.2695$$
$$SS_4 = SA_4 \div (a/b + \overline{7 - s}/2s) = 5.350 \div (19.2647 + 0.375) = 0.2724$$

1910 1st quar. $N = SA + 0.2382x = 4.500 + 0.2382(-1.5) = 4.1427$
2nd quar. $N = SA + 0.2390x = 4.575 + 0.2390(-1.5) = 4.2165$
3rd quar. $N = SA + 0.2695x = 5.225 + 0.2695(-1.5) = 4.8208$
4th quar. $N = SA + 0.2724x = 5.350 + 0.2724(-1.5) = 4.9414$

III. Normal (1st quar. 1911 = 4.1427 + 0.2382 = 4.38, etc.).

Quarter	1910	1911	1912	1913
January–March	4.14	4.38	4.62	4.86
April–June	4.22	4.46	4.69	4.93
July–September	4.82	5.09	5.36	5.63
October–December	4.94	5.21	5.49	5.76

The link-relative seasonal index.—The so-called link-relative method of finding the seasonal index was developed by Professor Persons in connection with the time series analysis studies carried on by the Harvard Committee of Economic Research, and published in the first volume of the *Review of Economic Statistics.* This method does not require the finding of the moving average, but begins the calculation of the seasonal by computing the link-relatives; that is, the ratios in percentages of each item, except the first, divided by the preceding item. This process in effect reduces each item to an index number of which the preceding item is the base. These index numbers are then combined into a seasonal index. The entire procedure may be outlined as follows:

(1) Linking: Divide the second item of the given time series by the first item to obtain the link-relative of the second item; the third by the second, to obtain the link-relative of the third; and so on, to the end of the series. The resulting link-relatives are index numbers having the preceding item as a base. They are tabulated like the seasonal percentages (cf. Table 12, p. 196).

(2) Averaging: Average by like periods (the average of the link-relatives for each January, then for each February, etc., or similarly by quarters for quarterly data). The median or the average of a few of the median items is used, thus eliminating the effect of extreme and irregular fluctuations.

(3) Chaining: Chain the averages thus obtained; that is, multiply the first by the second, the product thus obtained by the third, etc., thus reversing the calculation of the link-relatives. The base of the crude chain index thus obtained is 100 in the last period of the prior year. If no trend effect is present, the last item in the index will also be 100.

(4) Leveling: In case the last item in the chain index thus obtained is above or below 100, subtract (algebraically) from each item the fraction of the discrepancy corresponding to the given period of the year (as $\frac{1}{12}$ from January, $\frac{2}{12}$ from February, etc., or $\frac{1}{4}$ from the first quarter, etc.). The last item will now be 100, and the others will be changed according to their distance from the constant 100 of the prior year. In general, if the discrepancy is $d(d =$ last item of crude chain less 100), and the number of subdivisions is s, subtract from the successive items of the crude index, respectively, d/s, $2d/s$, $3d/s \ldots sd/s$. The result is a leveled index from which trend influence has presumably been removed. It is a little more precise to make this adjustment logarithmically, but if such accuracy is required the whole process of combining the link-relatives should preferably be put on a logarithmic basis.

(5) Centering: The leveled index may now be centered, so that its average is 100. This is preferably done by ratios rather than by additions, so as to keep the relative size of the items unchanged. Each item is therefore divided by the average of all the items. The results, expressed as percentages, constitute an index of seasonal variations.

The method of link-relatives, as thus described, is applied to interest rates in the United States by quarters, 1909–1913, in Example 64. It will be seen that the resulting seasonal index differs considerably from that obtained by the moving average method. With more adequate and more representative data, however, the two results would be fairly comparable. It is difficult to say which of the two processes is the better, but the moving average method seems to be somewhat more generally used. Since the moving average is itself of interest in time series analysis, and is useful in certain cases of projecting the trend, it is quite commonly computed; and after it has been computed, the seasonal based upon it is easily obtained.

Example 64.—Link-relative method of computing the seasonal index, applied to commercial interest rates by quarters, United States, 1909–1913. The link-relatives are obtained by dividing the second item by the first (3.9/3.8 = 102.6), the third item by the second (4.2/3.9 = 107.7), etc., to the end of the series (5.8/6.0 = 96.7). The average of two or three median items is then taken by quarters. These averages are chained by successive multiplication to obtain the crude chain. The chain is leveled by subtracting from the last item the difference between it and 100; from the third item, three-quarters of this difference; from the second item, one-half of this difference; and from the first item, one-fourth of this difference. The leveled chain is centered by dividing by its own average (or preferably this centering is applied to the logs, and antilogs taken). The resulting index, as indicated below, differs somewhat from that previously obtained by the moving average method (92.7; 92.6; 103.4; 111.3) but ordinarily the two methods parallel each other rather closely.

Data (Y)

Quarter	1909	1910	1911	1912	1913
January–March	3.8	4.7	4.1	3.9	5.3
April–June	3.9	4.8	3.6	4.2	5.7
July–September	4.2	5.6	4.2	5.1	6.0
October–December	5.5	5.3	4.3	6.0	5.8

Link-relatives (LR; given item as percentage of previous item)

Quarter	1909	1910	1911	1912	1913	Av. of md. items	Crude chain	Chain leveled	Index (centered)
Jan.–Mar.	.....	85.5	77.4	90.7	88.3	86.90	86.90	84.76	92.3
Apr.–June	102.6	102.1	87.8	107.7	107.5	104.07	90.44	86.16	93.8
July–Sept.	107.7	116.7	116.7	121.4	105.3	113.70	102.83	96.41	105.0
Oct.–Dec.	131.0	94.6	102.4	117.6	96.7	105.57	108.56	100.00	108.9
						410.24		4)367.33	400.0
								91.8325	

EXERCISES

1. What is the purpose of time series analysis? Outline the successive steps. Why is the method of differences applied to the logarithms of the data considered the best method in theory?

2. Each of the following series represents quarterly data for three consecutive years. Using the moving average method, compute approximately the seasonal index and the cycle. In each case where cyclic change is found, plot (1) the data and the normal, and (2) the cycle, in *AD* units. If no cyclic change is found, plot together data and trend.

(a) 80, 59, 104, 119; 84, 63, 116, 137; 97, 72, 127, 142.
(b) 92, 91, 102, 107; 90, 96, 103, 111; 94, 95, 107, 112.
(c) 88, 68, 104, 126; 87, 70, 113, 132; 94, 72, 114, 138.
(d) 90, 93, 91, 94; 98, 101, 99, 102; 106, 109, 107, 110.
(e) 6, 7, 12, 15; 14, 15, 20, 23; 22, 23, 28, 31.
(f) 90, 96, 168, 176; 162, 160, 264, 264; 234, 224, 360, 352.
(g) 48, 49, 55, 56; 56, 57, 63, 64; 64, 65, 71, 72.

3. Each of the following series represents quarterly data for three consecutive years. Using the link-relative method, compute approximately the seasonal index and the cycle. Chart each problem as in preceding exercise.

(a) 80, 59, 104, 119; 84, 63, 116, 137; 97, 72, 127, 142.
(b) 88.4, 92, 94.6, 93; 95.9, 100, 103.1, 101; 104, 108.1, 111.2, 108.7
(c) 209, 205, 211, 207; 201, 197, 203, 199; 193, 189, 195, 191.
(d) 208, 204, 210, 206; 200, 196, 202, 198; 192, 188, 194, 190.
(e) 192, 196, 190, 194; 200, 204, 198, 202; 208, 212, 206, 210.

4. Reduce to *AD* and σ units the following cycles of wholesale prices and interest rates.

Wholesale prices

Quarter	1909	1910	1911	1912	1913
1	−12	11	− 5	− 1	7
2	−10	5	−11	6	−1
3	− 4	2	− 5	7	−3
4	8	− 1	− 2	11	−2

Interest rates

Quarter					
1	− 3	13	− 6	−15	10
2	− 2	14	−19	−10	17
3	− 6	18	−15	− 2	10
4	12	3	−21	5	− 3

5. The following quarterly data represent certain classes of business failures in the Middle West, deflated for price changes, for the years designated. Using the moving average method, compute the cycle.

Quarter	1923	1924	1925	1926	1927	1928	1929
1	621	1691	1008	672	656	690	556
2	695	528	446	1399	1212	364	450
3	1253	1278	410	347	195	482	349
4	620	1115	606	473	468	651	363

6. The following monthly data represent bank debits deflated for price changes (5 ciphers omitted), for eight Iowa cities for the years indicated. Using the moving average method, compute the cycles. Quarterly totals, as given below, may be substituted.

Month	1923	1924	1925	1926	1927	1928	1929	1930
January......	1756	1606	1877	1783	1754	1572	1647	1833
February.....	1434	1445	1494	1481	1509	1432	1433	1543
March.......	1962	1745	1855	1855	1897	1749	1724	1785
April........	1734	1616	1762	1742	1712	1600	1699	1801
May.........	1728	1678	1609	1664	1654	1683	1681	1792
June.........	1871	1552	1626	1709	1719	1757	1646	1737
July.........	1651	1604	1685	1719	1674	1618	1721	1703
August.......	1552	1451	1540	1536	1568	1593	1694	1584
September....	1545	1563	1537	1587	1641	1584	1629	1615
October......	1766	1757	1821	1767	1745	1762	1864	1758
November....	1545	1471	1550	1591	1568	1535	1735	1508
December....	1662	1547	1705	1819	1651	1669	1725	1650
Jan.–Mar.....	5152	4796	5226	5119	5160	4753	4804	5161
Apr.–June....	5333	4846	4997	5115	5085	5040	5026	5330
July–Sept.....	4748	4618	4762	4842	4883	4795	5044	4902
Oct.–Dec.....	4973	4875	5076	5177	4964	4966	5324	4916

7. The following individual percentage cycles (data divided by statistical normal) are quoted from the *Annalist*, by months, 1929–1930. The average deviation as computed for prior years by the *Annalist* are also quoted, together with weights representing an estimate of the importance of each series relative to the measurement of the business cycle in the United States. Calculate the individual percentage cycle (d/AD) and make a composite of the five series using the given weights. For method and results see Example 60 and Chart 29.

1929	Pig iron	Freight car loadings	Electric power	Auto production	Cotton consumption
January.............	109.6	101.2	103.1	145.8	111.2
February............	108.7	104.5	102.1	142.8	107.7
March..............	108.4	101.2	100.5	142.7	107.9
April...............	110.4	107.5	104.0	141.8	110.7
May................	116.3	106.4	105.4	137.8	113.5
June................	123.1	105.6	101.8	144.9	104 8
July................	127.4	105.9	104.1	137.4	104.9
August..............	126.3	105.6	105.0	124.1	100.4
September..........	119.7	105.6	102.3	123.2	103.6
October.............	112.9	101.8	103.4	121.8	108.7
November...........	103.7	95.8	99.3	104.7	93.8
December...........	91.7	96.1	97.6	72.5	85.8

1930	Pig iron	Freight car loadings	Electric power	Auto production	Cotton consumption
January	89.9	95.7	98 6	99 9	92 9
February	96.0	96.3	96.3	106.3	86 5
March	95.0	92 7	94.7	98.1	84.2
April	95.3	96.5	97.5	100.9	90.8
May	95.5	94.0	95.2	94.9	76 7
June	95.9	91.7	93.9	89.5	76.7
July	87.3	90.5	94.5	71.9	75.2
August	84.3	88.6	91.1	57.9	67.9
September	78.2	85.8	90.9	61.6	72.2
October	68.1	83.9	88.6	47.3	72.8
November	60.6	80.1	85.3	64.3	72.0
December	53.5	80.0	84.7	86.0	71.4
Average deviation	20	5	4	22	9
Weight	10	20	10	10	15

8. Using a moving average method of time analysis, determine the seasonal, trend, and cycle elements in the following quarterly data for the United States.

(a) Polished plate glass production (100,000 sq. ft.).

Quarter	1923	1924	1925	1926	1927	1928	1929	1930
1	211	233	270	329	299	315	359	302
2	229	239	295	343	279	328	376	331
3	222	205	306	335	288	330	418	218
4	229	239	301	282	247	334	352	206

(b) Production of sole leather and thousands of backs, bends, and sides.

Quarter	1923	1924	1925	1926	1927	1928	1929	1930
1	1599	1268	1268	1056	1267	1275	1175	1297
2	1662	1128	1302	1115	1376	1388	1191	1329
3	1592	1182	1226	1150	1362	1408	1218	1337
4	1391	1304	1153	1217	1271	1308	1258	1227

9. Employing any of the accepted methods, prepare a time series analysis of the following quarterly combined index of industrial production, from *Survey of Current Business*, Year Book, 1932:

Quarter	1923	1924	1925	1926	1927	1928	1929	1930
1	102.00	102.00	106.33	107.67	110.00	109.33	120.67	106.00
2	106.67	90.00	102.33	107.00	109 67	109.33	125.00	103.67
3	100.67	87.67	100.67	108.33	104 33	110.33	121.67	91.00
4	97.67	98.00	106.00	108.33	100 67	114.00	108 33	83.67

10. Employing any of the accepted methods, prepare a time series analysis of the following quarterly combined index of business activity, from *Survey of Current Business*, Year Book, 1932:

Quarter	1923	1924	1925	1926	1927	1928	1929	1930
1	109.1	103 9	102.6	103.4	104.6	99.0	105.3	93.5
2	113.7	92.9	101.1	102.8	104.2	100.0	109.3	91.3
3	108.0	90.8	100 9	105.0	101.4	101.7	108.4	84.0
4	102.5	98.9	104.0	105.5	95.8	103.6	98.2	77.2

11. Prepare a time series analysis on the basis of the following data of freight car loadings as published by the Federal Reserve Board:

Quarter	1923	1924	1925	1926	1927	1928	1929	1930
1	90 67	93.33	94 67	96.33	99.00	94.33	97.33	90.00
2	100 67	92.67	100.33	104.33	103.00	100.67	107.00	95.00
3	107.33	101.33	109.67	114.33	109 67	111.00	115.67	96.67
4	100.67	103.00	106.33	111.00	101.00	107.33	103.00	85.67

12. (I) From the following quarterly seasonal percentages (Y/MA; a and b) for the years 1901–1904, calculate centered indexes of seasonal variation.

(a) Seasonal percentages

Quar.	1901	1902	1903	1904
1	....	40	80	60
2	....	100	120	80
3	80	60	100	....
4	120	140	160	....

(b) Seasonal percentages.

Quar.	1901	1902	1903	1904
1	....	70	60	80
2	....	80	100	90
3	140	130	150	
4	110	130	120	

(II) From the following seasonal differences (detrended data less its moving average) calculate centered indexes of seasonal variation.

(a) Seasonal differences.

Quar.	1901	1902	1903	1904	1905
1	..	−3	−10	1	−5
1	..	3	− 1	0	0
3	4	2	− 1	7	..
4	4	6	1	10	..

(b) Seasonal differences.

Quar.	1901	1902	1903	1904	1905
1	..	−5	5	−4	0
2	..	3	5	−1	6
3	10	3	8	4	
4	−5	−8	1	−3	

13. (a) Professors Reinhardt and Davies in "Principles and Methods of Sociology" give the cycle figures for the fluctuations in several social series. Reduce each of these cycles to average deviation cycles· and plot the results. The data given in the columns labeled "cycles" are deviations from a trend. (See pp. 621 to 623.)

(b) Various series of data should be secured and subjected to time series analysis. For example, the data for

(1) Automobile production may be found on p. 359 of the *Statistical Abstract of the United States*, 1932;

(2) Pig iron production may be found on p. 200 of the *Survey of Current Business*, Annual Supplement, 1932;

(3) Value of exports and imports may be found on p. 430, *Statistical Abstract of the United States*, 1932, etc.

ANSWERS

1. Consult various text-books (cf. bibliography).

2.

(a)	$a = 100.0$	$b = 9.5$	Seasonal index,	90.0; 65.0; 114.0; 131.0
(b)	$a = 100.0$	$b = 2\ 0$	" "	92.0; 94.9; 103.4; 109.5
(c)	$a = 100\ 5$	$b = 4.0$	" "	90.0; 70.0; 110.0; 129.0
(d)	$a = 100.0$	$b = 8.0$	" "	101.0; 103.0; 97.0; 99.0
(e)	$a = 18\ 0$	$b = 8.0$	" "	93.5; 89.2; 106.1; 111.2
(f)	$a = 212.5$	$b = 80\ 0$	" "	90.6; 79.5; 119 0; 110.9
(g)	$a = 60.0$	$b = 8.0$	" "	98.3; 96 8; 103.5; 101.7

3. (a) $a = 100\ 0$ $b = 9.5$ Seasonal index, 90.0; 63.5; 115.4; 131.0
(b) $a = 100\ 0$ $b = 8.0$ " " 99.0; 101 0; 102 0; 98.0
(c) $a = 200.0$ $b = -8.0$ " " 99.0; 98 0; 102.0; 101.0
(d) $a = 199.0$ $b = -8.0$ " " 99 0; 98.0; 102 0; 101.0
(e) $a = 201.0$ $b = 8\ 0$ " " 101.0; 102.0; 98 0; 99.0

4. *AD* and σ cycle of wholesale prices: $AD = 5.7$, $\sigma = 6.7823$.

Quarter	1909	1910	1911	1912	1913
1	−2.11	1.93	−0.88	−0.18	1.23
2	−1.75	0.88	−1.93	1.05	−0.18
3	−0.70	0 35	−0.88	1.23	−0.53
4	1.40	−0.18	−0.35	1.93	−0 35

Quarter	1909	1910	1911	1912	1913
1	−1.77	1 62	−0.74	−0.15	1 03
2	−1.47	0 74	−1.62	0.88	−0 15
3	−0.59	0.29	−0.74	1.03	−0 44
4	1.18	−0.15	−0.29	1.62	−0.29

AD and σ cycle of interest rates: $AD = 10.2$, $\sigma 11.84$.

Quarter	1909	1910	1911	1912	1913
1	−0.29	1.27	0 59	−1.47	0.98
2	−0 20	1.37	−1.86	−0 98	1.67
3	−0.59	1.76	−1.47	−0.20	0.98
4	1.18	0.29	−2.06	0.49	−0.29

Quarter	1909	1910	1911	1912	1913
1	−0.25	1.10	−0.51	−1.27	0 84
2	−0.17	1.18	−1.60	−0.84	1.44
3	−0.51	1.52	−1.27	−0.17	0.84
4	1.01	0.25	−1.77	0.42	−0.25

5. Seasonal index by quarters, 123.1; 90.8; 88.1; 98.0.
Trend equation: $T = 699.93 + 82.16x$.

6. Seasonal index by months, 103.4; 89.0; 109.3; 102.6; 101.8; 101.6; 100.7; 93.3; 95.3; 107.4; 93.9; 101.4.

Trend: $T = 1664.7 + 4.2888x$.

$AD = 2.954$.

AD cycle, January, 1923, $= +1.04$.

Seasonal index by quarters, 100.35; 101.18; 96.85; 101.62.

Trend: $T = 4994.25 + 12.95x$.

$AD = 2.54125$.

AD cycle, January, 1923, $= +1.51$.

7.

	1929	1930		1929	1930
January	+0.874	−0.579	July	+1.119	−1.726
February	+0.921	−0.703	August	+0.918	−2.282
March	+0.654	−1.048	September	+0.839	−2.373
April	+1.262	−0.577	October	+0.716	−2.740
May	+1.337	−1.222	November	−0.383	−3.061
June	+1.028	−1.448	December	−0.953	−3.008

8. (a) Polished plate glass production, time analysis. The following results may be taken as approximate:

Seasonal: Quarters

Quarters		Quarters	
1st	100.86	3rd	100.29
2nd	103.04	4th	95.80

$T = 289 + 13.03x.$

$AD = 11.59\%$ of normal.

Cycle, in average deviation units:

Quarter	1923	1924	1925	1926	1927	1928	1929	1930
1	−1.07	−0.70	0.10	1 50	0.18	0.23	1.06	−0.79
2	−0.70	−0.78	0 61	1.59	−0.68	0.32	1.21	−0.30
3	−0.82	−1.79	1.06	1.50	−0.28	0.53	2.49	−3.04
4	−0.33	−0.41	1.26	0.22	−1.22	0.96	1.09	−3.16

(b) Production of sole leather and thousands of backs, bends, and sides, time analysis. The following results may be taken as approximate:

Seasonal: Quarters

Quarters		Quarters	
1st	97.67	3rd	101.27
2nd	101.56	4th	99.68

$T = 1290.65 - 16.5x.$

$AD = 7.6\%$ of normal.

Cycle, in average deviation units:

Quarter	1923	1924	1925	1926	1927	1928	1929	1930
1	2.7	−0 4	−0.2	−2.2	0.1	0.3	−0.5	0.9
2	2.8	−2.2	−0.4	−2.1	0 8	1.0	−0.8	0.8
3	2.2	−1.6	−1.0	−1.6	0.7	1.3	−0.5	1.0
4	0.5	−0.2	−1.5	−0.7	0.0	0.5	0.2	0.0

9. Moving average, first and last: 101.75, 99.17.
Corrected seasonal index: 101.7, 101.5, 98.0, 98.8.
a: 105.0; b: 1.2.
Trend, first 100.3, last 109.7.
AD: 5.42.
d/AD: first −0.015, last −4.205.

10. Moving average, first and last: 107.7, 89.1.
Corrected seasonal index: 100.0, 100.3, 100.2, 99.5.
a: 100.706; b: 1.3725.
Trend, first and last: 106.0, 95.4.
AD: 4.872.
d/AD, first and last: 0.5971, −3.8281.

11. Compare percentage cycle with *Annalist* cycle of car loadings as published monthly (e.g., June 19, 1931, p. 1107).

12. (I) $a = 63, 105, 84, 147.$ $b = 67, 86, 133, 114.$
(II) $a = -6, -1, 2, 4.$ $b = -3, 3, 5, -5.$

CHAPTER VIII

CORRELATION

IN the preceding chapter it was shown that time series analysis results in the isolating of a cycle which may be expressed in average or standard deviation units, thus making it mathematically comparable with other cycles similarly derived. The question often arises as to the degree of resemblance between two such cycles; for example, when manufacturing is active, is employment generally high? Such a comparison may be made by working out the cycles of manufacturing production and employment from adequate data, plotting them to the same scale, and observing whether they have a tendency to coincide. If so, they are said to be more or less correlated, and since they tend to agree they are said to be positively correlated. In the same way, cycles of the corn crop in the United States might be plotted by years, and compared with cycles of the price for the same years. These cycles would tend to move in opposite directions, and therefore would be said to be inversely, or negatively, correlated. Similar comparisons may be made for other data than time series by comparing deviations from some average or standard taken as the normal; thus we might correlate the percentage of foreign born in various states with urbanization, or the death rate with the development of systems of public health. It will be observed that such correlations do not necessarily imply cause and effect, although such a relation may often be inferred.

The degree of correlation.—It is evident from a comparison of cycles or other deviations that correlation is usually a matter of degree. It would be very seldom, indeed, that two cycles would be exactly alike, positively or negatively. When manufacturing increases, employment generally increases, but not always in quite the same degree, since the use of machinery, the length of the working day, and other factors will produce variations. Sometimes the other variables may be so significant as to obscure an expected correlation.

Often it is desirable in the study of social interrelations to express definitely the degree of correlation as it appears on the basis of a given study; that is, to state it as a coefficient. This is done mathematically on the basis of a scale which ranges from +1.00, expressing complete

likeness of the cycles, to −1.00, expressing inverse likeness of the cycles. Both of these extreme cases represent complete correlation, one of the positive and the other of the negative type. In the former case the cycles coincide; in the second case they would coincide if the signs of one series were reversed. The calculation of the coefficient of correlation involves a measure of the degree of similarity of the two cycles in question, expressing how close they come to complete positive or negative correlation. The scale that is used is mathematically arbi-

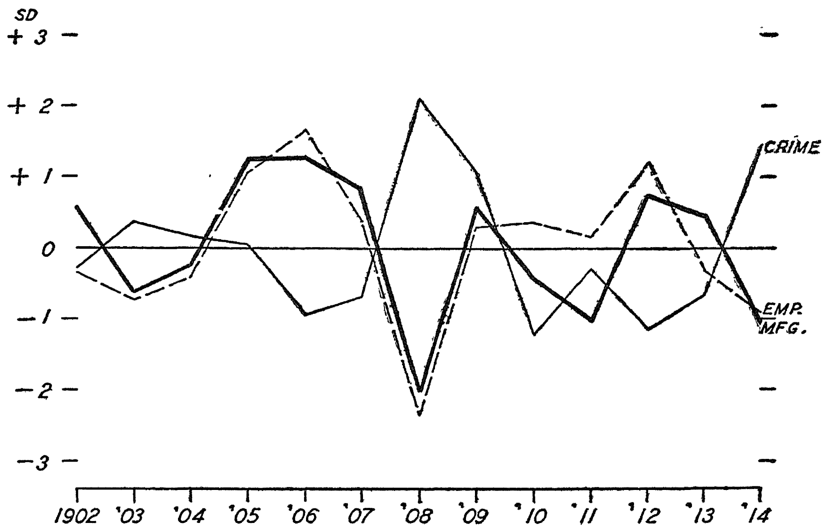


CHART 32

Standard deviation cycles of manufacturing (mfg.), employment (emp.), and crime rates (crime) in the state of New York, by years, 1902–1914. The correlation of employment and manufacturing is positive ($r = 0.82$); the correlation of employment and crime is negative ($r = -0.75$). The data are insufficient to yield a valid coefficient expressing a general rule, but are presented to introduce in elementary form the nature of positive and negative correlation.

trary; for example, a coefficient of 0.80 simply indicates a marked degree of similarity, and a coefficient of −0.80 indicates a marked approach toward reversed similarity. The coefficient will be given a more precise mathematical meaning later, but its real significance must be developed chiefly out of experience with it, just as a measurement by a thermometer or a yardstick comes to have significance through experience.

Examples of correlation.—In general, the nature of the correlation of two cycles may be represented by the data of Example 65 which are

plotted in Chart 32. The chart represents the standard deviation cycles of manufacturing, employment, and crime in New York State, by years from 1902 to 1914. It may be seen from the chart that employment and manufacturing have a marked tendency to move together, although in 1910 and 1911 employment remained above normal while manufacturing was considerably below normal. At the same time the fluctuations of crime, as represented by the data at hand, ran for the most part opposite to the cycles of employment. If employment and manufacturing cycles had exactly coincided as they nearly did from 1903 to 1909, the correlation coefficient (r) would have been $r = 1.00$. But the various divergences pull down the coefficient so that it becomes $r = 0.82$. The cycles of crime and employment fall considerably short of being exact opposites, and give a correlation of $r = -0.75$. The method of computing these correlations will be considered later; the chart will, however, serve the purpose of suggesting the significance of the correlation coefficient.

Example 65.—Standard deviation cycles of manufacturing, employment, and crime, New York State, 1902–1914. It will be noted that the manufacturing and employment cycles are much alike; that is, they are positively correlated, whereas employment and crime tend to be negatively correlated. The degree of correlation is 0.82 in the first case, and −0.75 in the second. Methods of computing the degree of correlation will be considered later. The cycles are plotted in Chart 32.

Year	(1) Manufacturing	(2) Employment	(3) Crime
1902	0.59 (σ)	−0.37 (σ)	−0.29 (σ)
1903	−0.61	−0.75	0.40
1904	−0.23	−0.41	0.18
1905	1.25	1.05	0.07
1906	1.29	1.66	−0.96
1907	0.82	0.37	−0.71
1908	−2.07	−2.35	2.13
1909	0.55	0.31	1.07
1910	−0.48	0.37	−1.23
1911	−1.05	0.16	−0.30
1912	0.75	1.20	−1.17
1913	0.46	−0.34	−0.67
1914	−1.27	−0.90	1.48

Allowing for lag.—When two time series are compared, as by graphing their cycles in average deviation units on the same chart, it will sometimes be found that there is an obvious relation between the two sets of cycles, but that one tends to precede and the other to lag. For example, during the two decades before the World War, the cycle of stock prices, as derived from data of the New York Stock Exchange, tended to precede general business activity by a

few months, whereas the cycle of interest rates tended to lag after business activity (cf. p. 213).

When such precedence or lag is clearly established, the correlation may be made by dating the lagging series uniformly earlier so that the cycles appear to coincide or contrast most markedly. This may conveniently be done by plotting one of the series to like scales on transparent paper, superimposing it on the other, and shifting it back and forth until the place of greatest concurrence or contrast appears to be reached. If greater accuracy is required, the correlation may be computed at several points about the point of greatest apparent correlation and the degree of lag selected that gives the largest positive or negative coefficient.

The measure of correlation.—In order to introduce the mathematical measurement of correlation, that is, the computation of the coefficient, it is desirable first to plot the data as a so-called scatter diagram

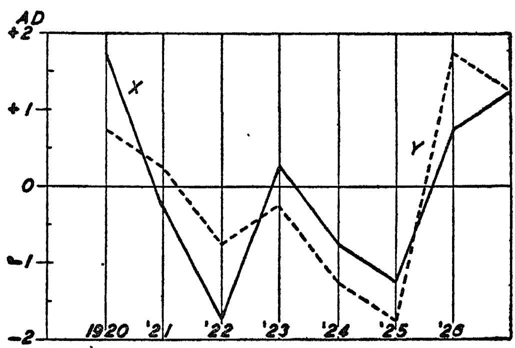


CHART 33

Assumed average deviations cycles, 1920–1927, as follows:

x: +1.75; −0.25; −1.75; +0.25; −0.75; −1.25; +0.75; +1.25
y: +0.75; +0.25; −0.75; −0.25; −1.25; −1.75; +1.75; +1.25

The cycles indicate a fair degree of positive correlation, but the data are inadequate to give valid results. They are replotted in a scatter diagram in Chart 33a.

rather than as cycles. This procedure will make possible a more accurate analysis of the paired deviations, and will also make possible the presentation of correlated deviations for other than time series. The scatter diagram of Chart 33a is a replotting of the cycles presented in Chart 33. It represents two paired x and y deviations as one point measured upon the x- and y-scales. The data for 1920, for example, in terms of the average deviation cycles, are: $x = 1.75$ and $y = 0.75$;

the point representing the year 1920 is therefore plotted as a coordinate of these positions on the x- and y-scales, respectively. In the same way each point represents two paired deviations of the cycles. Such a chart, though not visually as informing as the cycle chart, will afford

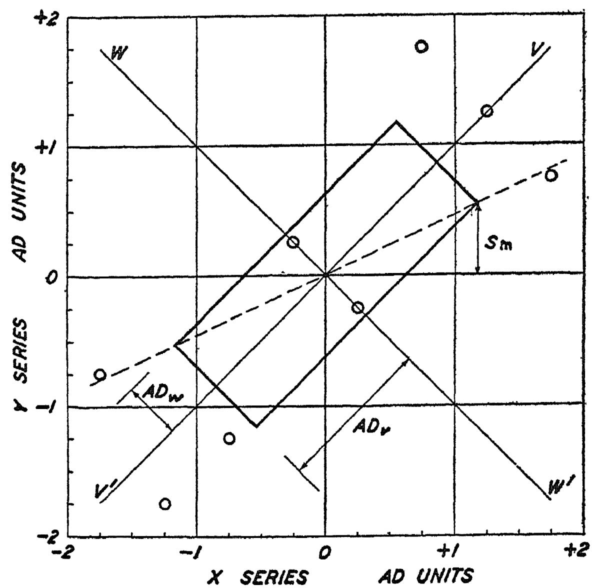


CHART 33a

Scatter diagram of the average deviation cycles plotted in Chart 33. Each pair of deviations is here plotted by one point read on the coordinate scales x and y (AD units). The degree of positive correlation is indicated by the tendency of the data to cluster about the positive diagonal, $V'V$. The average deviation of the plotted points, measured perpendicularly to the negative diagonal WW', is indicated as AD_v, and the average deviation from the positive diagonal is indicated as AD_w. These two average deviations determine the central rectangle, which would become a square if there were no correlation, and would narrow to the diagonal of +1 or −1 if there was perfect correlation. The relative narrowness of this rectangle as measured by $(AD_v - AD_w)/\sqrt{2}$ is the coefficient of similarity (Sm). If the same data had been measured in standard deviation units throughout, the coefficient of correlation (r) would be $(\sigma_v^2 - \sigma_w^2)/2$, which would also be a function of the central rectangle in terms of the second moments. Graphically, for normally distributed data, r would be the product of Sm and its abscissa.

a better measure of the mathematical degree of correlation, as the following considerations will show.

It will readily be seen that if the deviations in the two cycles are exactly alike ($r = + 1.00$), for example, if when $x = + 0.25$, $y = + 0.25$, and when $x = - 1.75$, $y = - 1.75$, the plotted points will form a positive diagonal line ($v'v$) from left to right upward across the chart. If, however, the cycles are reversed ($r = - 1.00$), that is, for example, if

when $x = 0.25$, $y = -0.25$, and when $x = -0.75$, $y = +0.75$, the plotted points will form a negative diagonal (ww') downward from left to right across the chart. If, on the other hand, there is no correlation, and the paired deviations concur as if by chance, then the plotted points will be scattered in a haphazard way as if they had been located by chance. But if there is a moderate degree of correlation, either positive or negative, then the points will tend to cluster about one or the other of the diagonals, and the degree of correlation will be sug-

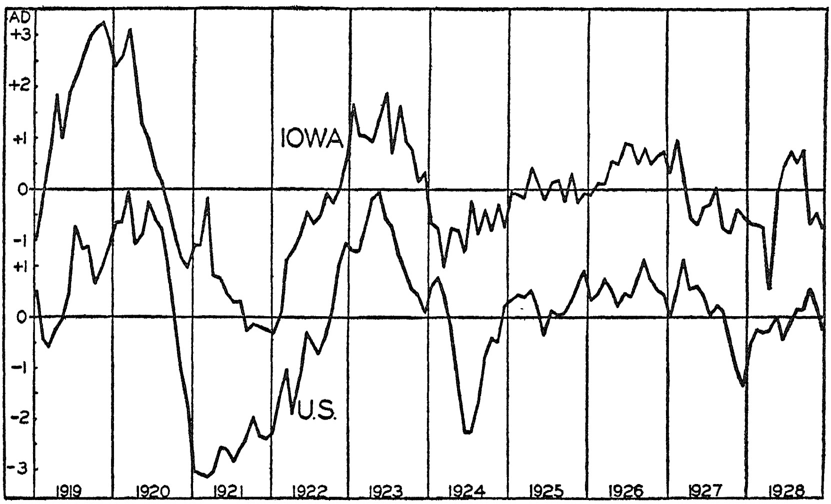


CHART 34

Average deviation cycles of business activity in Iowa and the nation, by months, 1919–1928. The data are replotted in the form of a scatter diagram in Chart 34*a*. The degree of correlation is expressed by the coefficients $Sm = 0.57$ and $r = 0.75$; r computed from Sm is 0.74 (cf. Example 66 for method).

gested by their closeness to this diagonal. The degree of correlation may therefore be visualized by the arrangement of the plotted points (cf. Charts 34 and 34*a*).

The degree of correlation may be mathematically computed as follows: If the average deviation from a diagonal is measured by drawing perpendiculars from the points to the diagonal and averaging these deviations, the two resulting average deviations thus found from each diagonal will furnish a numerical basis for the coefficient of correlation. If the two average deviations are alike, there is no correlation; if one of them is zero there is perfect correlation of one type or the other. The degree of correlation may therefore be expressed as the difference between these two average deviations, divided by $\sqrt{2}$ to

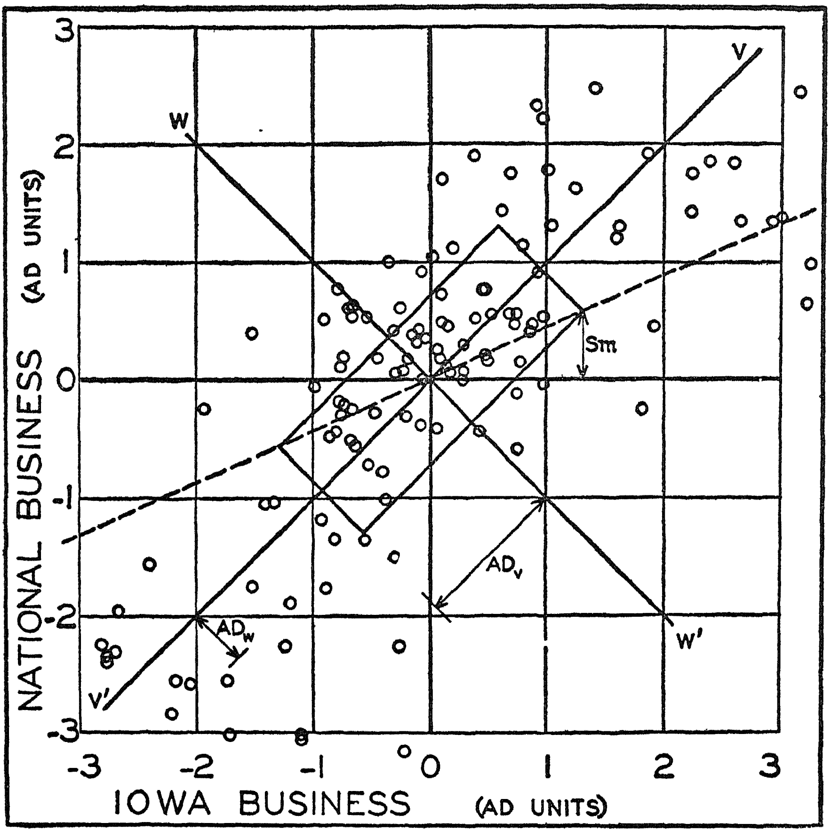


CHART 34*a*

Squared scatter diagram in average deviation units of cycles of business activity in Iowa and the United States (cf. Chart 34). The central rectangle measures the average deviations of the plotted points from the positive diagonal $V'V$, and the negative diagonal WW', respectively. The coefficient of similarity (*Sm*) is a function of this rectangle, as indicated graphically; and the coefficient of correlation in a similar chart constructed on a standard deviation basis throughout would be the line *Sm* times its abscissa. With normal distributions, *Sm* and *r* might be measured graphically on either an *AD* or a σ chart since the proportions would be alike in each case.

make it range from $+1$ to -1. The coefficient thus found is not the standard measure of correlation, however, but is a useful measure usually called the coefficient of similarity (*Sm*). Hence,

$$Sm = (AD_v - AD_w) \div \sqrt{2}$$

where AD_v is the average deviation of the plotted points from the negative diagonal (sloping down from left to right), and AD_w is the average deviation of the plotted points from the positive diagonal (sloping upward

from left to right). If, on the other hand, the standard deviation is used throughout in place of the average deviation, the standard coefficient of correlation (r) may be obtained on much the same principle, as

$$r = (\sigma_v^2 - \sigma_w^2) \div 2$$

Computing the coefficient.—The formulas just given, however, are not in a convenient form for use in computing Sm and r. They are given here merely to suggest the nature of the measurement of correlation on the basis of the rectangle shown in the squared scatter diagram, that is, the scatter diagram in σ or AD units. As actually applied to the correlation of average deviation or standard deviation cycles, they are changed algebraically into more convenient form as follows:

$$Sm = \Sigma s/n$$

where s is the numerically smaller of each pair of deviations in the AD cycles, written with the sign of the correlation (positive if the signs are alike and negative if the signs are unlike). Also,

$$r = \Sigma x_\sigma y_\sigma / n$$

where x_σ and y_σ represent paired items in the standard deviation cycles, respectively. The process is illustrated in Example 66.*

Example 66.—Correlation of average deviation (I) and standard deviation cycles (II) of manufactures in the United States and employment in New York State, by years, 1902–1914. The average deviation measure is $Sm = \Sigma s/n$, where s is the

* For the sake of brevity, the examples of correlation used in this chapter are based on insufficient data. It is difficult to determine the number of items necessary to establish a correlation, but in general twenty or twenty-five would be regarded as a bare minimum. If the items represent independent observations, the "probable error" is often applied. The probable error of r is $PE = 0.6745\ (1 - r^2) \div \sqrt{n}$. This is the quartile deviation of the normal curve theoretically formed by successive r's obtained on the basis of drawings of n paired items from a very large number of paired items correlated to the degree indicated by r. This measure may be misleading, however, particularly when n is small, in which case the reliability of the correlation can be estimated only by the use of complex tables (cf. articles by E. S. Pearson, H. L. Reitz, and P. R. Rider in the *Journal of the American Statistical Association*, June, 1931). Theoretical measures of probable validity generally assume independent observations; that is, items which are derived without reference to the preceding items. But in the business cycle each measure is an outgrowth of the preceding measure, and in geographic studies each district is more or less continuous with adjacent districts, hence the observations are not independent in the statistical sense. It is therefore very difficult to estimate the validity of a correlation in social data beyond that which may be derived from experience and from the support of successive studies. Hence, it is best not to attempt to deduce a general rule from a single correlation, but to use it merely tentatively as evidence to be weighed in accordance with the nature of the problem.

numerically smaller of the correlatives with the sign of correlation (like signs indicate positive correlation and unlike signs negative). The standard deviation measure is $r = \Sigma x_\sigma y_\sigma / n$, where x_σ and y_σ indicate the paired items in the standard deviation cycles, respectively. r is normally larger than Sm except when there is no correlation or perfect correlation. In normal distributions, $r^2 = 2Sm^2 - Sm^4$. If this formula is applied above, $r = (2 \times 0.5685^2 - 0.5685^4)^{1/2} = 0.736$, which may be taken as the value of r on the assumption of a smoothing of the data sufficient to remove in part the distorting effects of the second moment on irregular items.

	I. Average deviation cycles			II. Standard deviation cycles		
Year	Mfg.	Emp.	Smaller	Mfg.	Emp.	Product
1902	0.67	−0.47	−0.47	0.59	−0.37	− 0.2183
1903	−0.69	−0.95	0.69	−0.61	−0.75	0.4575
1904	−0.26	−0.52	0.26	−0.23	−0.41	0.0943
1905	1.42	1.33	1.33	1.25	1.05	1.3125
1906	1 47	2.11	1.47	1.29	1.66	2.1414
1907	0 93	0.47	0.47	0.82	0.37	0.3034
1908	−2.36	−2.98	2.36	−2.07	−2.35	4 8645
1909	0.63	0.39	0.39	0 55	0.31	0.1705
1910	−0 55	0.47	−0.47	−0 48	0.37	− 0.1776
1911	−1 20	0.20	−0.20	−1 05	0.16	− 0.1680
1912	0.85	1.52	0 85	0.75	1.20	0.9000
1913	0.52	−0.43	−0.43	0.46	−0.34	− 0.1564
1914	−1.45	−1.14	1 14	−1.27	−0.90	1.1430
			13) +7 39			13) +10 6668
			Sm = 0.5685			r = 0.8205

Comparison of Sm and r.—It will be seen by Example 66 that the coefficient of similarity which measures correlation on an average deviation basis is smaller than the coefficient of correlation which employs the standard deviation. This is normally to be expected, owing to the nature of the two measurements. It arises from the fact that in effect two different scales are used, although the scales coincide at the extremes (+1.00 and −1.00) and at the zero point (cf. Chart 35). It is somewhat as if instruments with different scales were used in physical measurements. The two scales are interchangeable, however, provided that the data are adequate and normally distributed. The relationship may be expressed as follows:

$$r^2 = 2Sm^2 - Sm^4$$

In a correlation employing inadequate data such as Example 66, this relationship may not be realized (r derived from Sm in Example 66 equals 0.74). But with more complete data the relationship will generally be rather close (cf. Charts 34 and 34a). Hence it is possible to compute the coefficient of correlation (r) by means of the coefficient of similarity (Sm). This process has some advantage in the case of

irregularly distributed data, in that it gives less weight to erratic items which are overemphasized by the standard deviation.* In general,

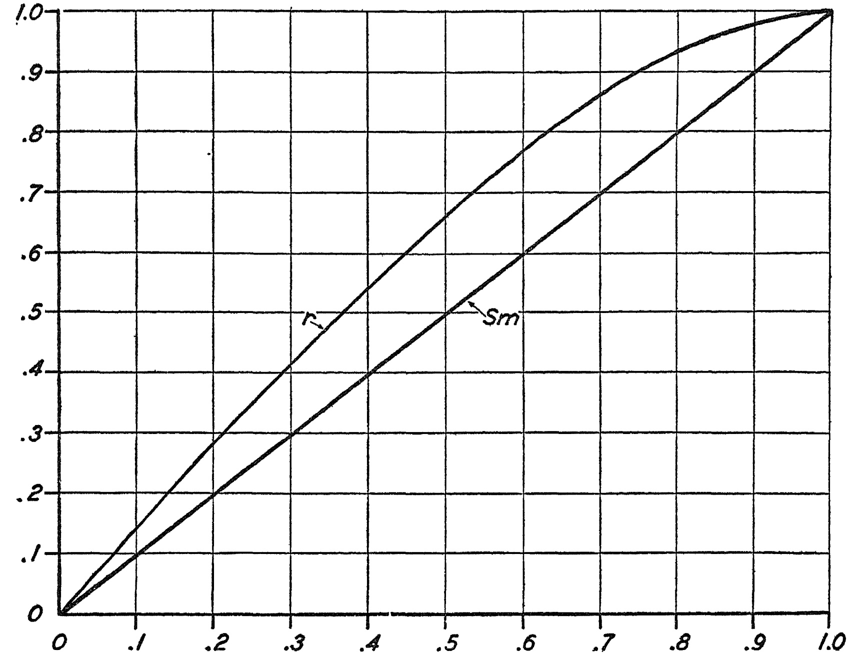


CHART 35

Comparison of values of the coefficient of similarity (Sm) and the Pearsonian coefficient of correlation (r) when computed from ample data normally distributed ($r^2 = 2Sm^2 - Sm^4$). If any given value of r is read on the vertical (Y) scale, the corresponding value of Sm may be found at the coordinate position in the horizontal (X) scale as read from the line r. For example, if $r = 0.8$, a horizontal line extended from 0.8 on the Y-scale will intersect the r line at a point directly above $Sm = 0.63$ on the Y-scale. The reading may be reversed to find the value of r corresponding to a given Sm. The line Sm reads the same on each scale, and shows the magnitude of Sm as compared with r on any ordinate from 0 to 1.0. The chart may be used for either positive or negative values.

* In correlations where a high Pearsonian coefficient (r) is obtained as a result of one or two pairs of extreme items, the coefficient of similarity is likely to prove a better measure of correlation than the Pearson r, because it tends to discount the effect of the extreme items. But if the distribution tends to cluster abnormally about the mode without having any extreme items, the coefficient of similarity may exaggerate the degree of correlation as compared with the Pearson r. In general, Sm is less sensitive to irregularities in the data than r. In changing Sm to its corresponding r, a second method may be employed which may sometimes neutralize this exaggeration: namely, $r = Sm(2 - \Sigma's'/n)$ where $\Sigma's'$ means the sum of the smaller correlatives in average deviation units taken without regard to the sign. With normal distributions having an adequate number of items, this formula will give practically the same result as $r^2 = 2Sm^2 - Sm^4$. When the results differ significantly, the smaller may generally be preferred. But in any case, coefficients of correlation based on irregular distributions are difficult to interpret and should not be given full weight.

however, the coefficient of similarity is likely to be used in less formal correlations which are not expected to yield a definite rule by which prediction may be made from one set of data to another.

General methods of finding *Sm* and *r*.—Whether the data to be correlated are time series or not, the logical first step in correlation is always the finding of the correlative sets of deviations. If the data are not time series, but represent measurements of certain conditions in a group of units such as cities or states, the deviations are taken from the mean. In any case, it is desirable, for the sake of brevity, to find a method which will eliminate either the reduction of each separate item to deviations, or the division of each separate deviation by the average or standard deviation of the series, or one which will eliminate both of these steps. These more general methods will now be considered, first for untabulated data and then for tabulated data.

Computing *Sm*, untabulated data.—The general method of calculating the coefficient of similarity in untabulated data is illustrated in Example 67, I. If the deviations have been taken from a normal, either a trend or a mean, the coefficient of similarity may be found without dividing each deviation by its appropriate average deviation, as follows: (1) Find the average deviation of each series, X and Y, and take the ratio (R) of the larger to the smaller (in this case, $R = AD_y/AD_x$). (2) Multiply the series of deviations having the smaller average deviation by the ratio just found. (3) Select the smaller of each pair of deviations, giving it its correlation sign (like signs, plus; unlike, minus), and total. (4) Divide this total by n times the larger average deviation. The result is the coefficient of similarity, which with adequate data may be reduced to an approximation of r as previously described ($r^2 = 2Sm^2 - Sm^4$). It will readily be seen that this process is the equivalent of that previously described, since the multiplying of one series by the ratio of the two average deviations equalizes the average deviations of the two series, and both average deviations are taken out in the final division.

If the deviations are taken from the mean rather than from the trend, a still shorter method may be developed (cf. Example 67, II), though this method is hardly worth mastering unless it is to be applied to laboratory routine. In this case the deviations are merely expressed as $x = X - AM_x$; and $y = Y - AM_y$, and the totals are divided by the respective average deviations. The smaller deviations, however, cannot be selected directly, but may be readily determined by a scatter diagram of the data (cf. Chart 36). The data are plotted or roughly located as Y against X, and the respective means are indicated on the scales together with their average deviations above and below the means.

The average deviation lines are then drawn forming a rectangle in the central portion of the chart, and diagonals are drawn through this rectangle. In such a chart the points falling in the upper and lower quadrants formed by the diagonals will indicate X items, and on the left and right quadrants Y items.* The deviations indicated above the negative diagonal (upper right half of chart) are written with a plus sign,

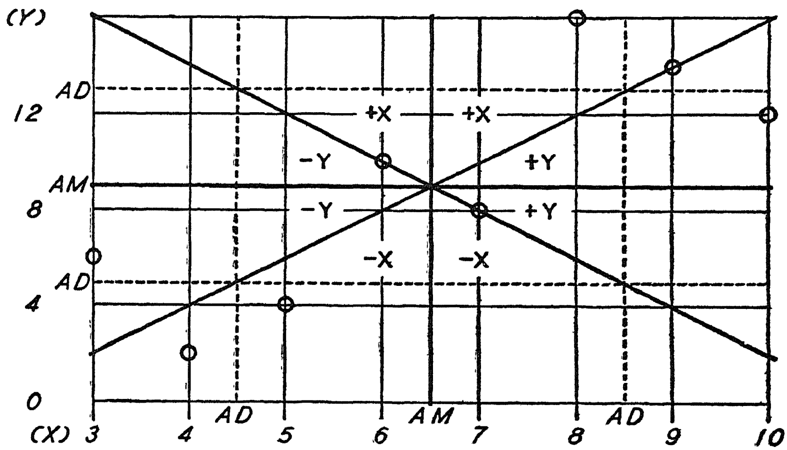


CHART 36

Scatter diagram for use in general method of finding coefficient of similarity in untabulated data measured from the means. The X and Y data with their means, plus and minus their respective average deviations, are plotted. Axes are drawn at the means, and lines parallel to them at the average deviation points above and below, and left and right of the means, respectively. Diagonals are drawn through the central rectangle formed by the average deviation lines. The deviations are expressed as each given item less the mean of its series ($X - AM_x$ and $Y - AM_y$). Select X-items of each pair plotted in the upper and lower quadrants formed by the diagonals, and Y-items of each pair falling in the left and right quadrants formed by the diagonals, and express their deviations as $x = X - AM_x$ and $y = Y - AM_y$, respectively. Items falling on a diagonal are preferably taken as if above the diagonal. Prefix minus signs to all expressed deviations ($X - AM_x$ and $Y - AM_y$) falling below the negative diagonal, that is, in the lower left-hand half of the chart, as indicated by the signs of X and Y in the central portion of the chart (these are not the actual correlation signs, but will become so when applied to the deviations as expressed). The coefficient of similarity may then be found as indicated in Example 67.

* When a point falls exactly on a diagonal, it may be classified as on either side of it, without prejudice to the result. But if it is classified as above the diagonal, it will often conveniently avoid a negative sign. If the diagonals of the chart are carefully drawn on a reasonably large scale, there can scarcely be an appreciable error in locating a point falling very close to a diagonal, even if it is classified on the wrong side. However, in such cases it is possible to determine the classification exactly by computing x/AD_x and y/AD_y and thus determining the smaller. In general, it is not necessary to plot the points on the scatter diagram; it is sufficient merely to determine the X or Y and plus or minus zone.

and those below this diagonal are written with a minus sign. This arrangement of signs, as indicated on the chart, does not conform to the true correlation signs, but when applied to the deviations, as here expressed, it will result in giving them their appropriate signs. The deviations are tabulated in X and Y columns, each of which is totaled, and divided by its average deviation. The grand total is then taken as Σs, which is divided by n to find the coefficient of similarity. The result may be changed to an approximation of the coefficient of correlation as before described ($r^2 = 2Sm^2 - Sm^4$).

Example 67.—General method of computing the coefficient of similarity in untabulated data. (I) Computed from the deviations taken from a normal consisting of either an average or a trend. The deviations ($x = X - AM_x$ and $y = Y - AM_y$) are found (the slope of the trend by the method of semi-averages is zero, hence the trend is taken as identical with the mean). The average deviations are computed, and the larger is divided by the smaller; in this case $R = AD_y/AD_x$. The smaller deviations (x) are multiplied by R to make both average deviations identical. The numerically smaller (s) of the paired deviations thus equated is then written with the correlation sign (like signs plus, and unlike minus), and Σs is divided by n times the larger AD to obtain Sm. (II) When deviations are taken from the mean, Sm may be found from X and Y as follows: Draw up a scatter diagram of Y on X as in Chart 36. Select as the smaller deviation of each pair the $X - AM_x$ or $Y - AM_y$ indicated by the position of the plotted points, giving it the sign indicated (cf. explanation accompanying chart). These deviations are written in two columns x and y; each is totaled and the sum divided by its appropriate average deviation which is obtained from the X and Y series as indicated in II. The grand total is Σs, which, divided by n, gives $Sm = 0.625$.

I. Computation from deviations x and y.

Year	X	Y	x	y	xR	s (of y and xR)
1920	10	12	3.5	3	7	3
1921	6	10	−0.5	1	−1	−1
1922	3	6	−3.5	−3	−7	3
1923	7	8	0.5	−1	1	−1
1924	5	4	−1.5	−5	−3	3
1925	4	2	−2 5	−7	−5	5
1926	8	16	1.5	7	3	3
1927	9	14	2 5	5	5	5
	$\Sigma = 52$	$\Sigma = 72$	$'\Sigma' = 16.0$	$'\Sigma' = 32$	$'\Sigma' = 32$	$4 \times 8\overline{)20}$

$AM_x = 6.5$; $AM_y = 9.0$; $AD_x = 2.0$; $AD_y = 4.0$; $AD = 4.0$; $Sm = 0.625$.

R = larger AD ÷ smaller AD = $4.0/2.0 = 2$.

$Sm = \Sigma s$ (in larger AD units) ÷ n times larger AD = $20 \div (4 \times 8) = 0.625$.

II. Computations from X and Y by use of scatter chart.

Year	X	Y	x-deviations (smaller)	y-deviations (smaller)
1920	10	12	....	(12 − 9)
1921	6	10	(6 − 6 5)	
1922	3	6	....	−(6 − 9)
1923	7	8	−(7 − 6.5)	
1924	5	4	−(5 − 6.5)	
1925	4	2	−(4 − 6.5)	
1926	8	16	(8 − 6.5)	
1927	9	14	(9 − 6 5)	
Totals.....	52	72	$(AD_x)2\overline{)7-0}$	$(AD_y)\ 4\overline{)6-0}$
AM..... ..	6.5	9	3.5	1.5
AD.......	2	4		

$\Sigma s = 3\ 5 + 1.5 = 5.$
$Sm = \Sigma s/n = \frac{5}{8} = 0.625.$

"Larger than mean" less "Smaller than mean":

$$AD_x = [(10 + 7 + 8 + 9) - (6 + 3 + 5 + 4)] \div 8 = 2$$
$$AD_y = [(12 + 10 + 16 + 14) - (6 + 8 + 4 + 2)] \div 8 = 4$$

If necessary, the number of items in the "larger than mean" and "smaller than mean" groups is equalized by inserting the mean the requisite number of times.

The computation of r, untabulated data.—It has already been stated that the standard coefficient of correlation r is expressed by the formula

$$r = \Sigma x_\sigma y_\sigma \div n = \Sigma xy \div (n\sigma_x\sigma_y)$$

in which x and y are the correlative data (X and Y) expressed in deviations, σ_x and σ_y are the respective standard deviations, and n is the number of items in each of the correlative series. The deviations, x and y, may be taken from either the mean or a trend, according to the nature of the problem. The process requires the finding of the means, the deviations from the means, and the standard deviations of the two series. The formula may then be applied as given (cf. Example 68). If the standard deviations themselves are not of interest as measures of dispersion, the process may be abbreviated somewhat by writing the formula as:

$$r = \Sigma xy \div (\Sigma x^2 \Sigma y^2)^{1/2}$$

or

$$r = \Sigma xy \div (n\sqrt{\sigma_x{}^2\sigma_y{}^2})$$

in which case the finding of the standard deviations is omitted.

When the data consist of comparatively small numbers, and the deviations are taken from the mean rather than from a trend, the calculation of r may be based directly upon the X and Y series rather than upon the deviations themselves, by a transformation of the formula as follows:

$$r = (\Sigma XY - nM_xM_y) \div [(\Sigma X^2 - nM_x^2)(\Sigma Y^2 - nM_y^2)]^{1/2}$$

where X and Y represent the correlative data, and M_x and M_y are the respective arithmetic means. This formula may also be applied to any readjustment of the x and y series as deviations from an arbitrary origin. Such deviations may be substituted for X and Y, respectively, and the formula directly applied to these deviations. As will be seen from Example 68, II, the process requires the finding of the averages of the X and Y series, together with the squares of these averages. The expressions ΣX^2, ΣY^2, and ΣXY are calculated from the data. The magnitudes thus found may then be substituted in the equation for r. One disadvantage of the method is that it will be found necessary to carry the calculation to several significant figures, since the subtractions indicated in the formula often leave comparatively small differences.

Example 68.—General method of computing the coefficient of correlation (r) in untabulated data. The abbreviated data are mean annual temperatures (X) in degrees Centigrade and adjusted annual death rates per 1000 population for certain European countries in pre-war years. (I) Computation of r from the deviations taken from the normal consisting of either an average or a trend. In this example, which is not a time series, the deviations (x and y) are taken from the mean. The computation makes use of the formula $r = \Sigma xy \div (n\sigma_x\sigma_y)$, consequently σ_x and σ_y are first found, the products xy are then written, and the formula computed. If the σ's are not of interest, the computation may be shortened a little by using the formula $r = \Sigma xy \div (\Sigma x^2 \Sigma y^2)^{1/2}$. (II) When the deviations are taken from the mean, the computation may be based directly upon X and Y, although this has the disadvantage of requiring a considerable number of significant figures. The formulas, as indicated below, are obtained algebraically from the second formula just given. The data are, of course, too incomplete to give a valid coefficient.

I. Computation of r from deviations (x and y) from the mean taken as normal.

Country	X	Y	x	x^2	y	y^2	xy
Denmark....	7.2	17.7	−1.4	1.96	−1.7	2.89	2.38
France......	11.7	21.1	3.1	9.61	1.7	2.89	5.27
Holland.....	8.7	19.2	0.1	0.01	−0.2	0.04	−0.02
Italy.......	15.2	22.5	6.6	43.56	3.1	9.61	20.46
Norway.....	3.8	17.8	−4.8	23 04	−1.6	2.56	7.68
Sweden.....	5.1	18.0	−3.5	12.25	−1.4	1.96	4.90
Switzerland..	8.6	19 7	0	0	0.3	0.09	0
	7)60.3	7)136.0	0.1	7)90.43	0.2	7)20 04	40.67
	8.6	19.4		σ^2 = 12.92		σ^2 = 2.86	
				σ = 3.59		σ = 1.69	

$r = \Sigma xy/n\sigma_x\sigma_y = 40.67/(7 \times 3.59 \times 1.69) = 0.96,$

or $r = \Sigma xy/(\Sigma x^2\Sigma y^2)^{1/2} = 40.67/(90.43 \times 20.04)^{1/2} = 0.96.$

II. Computation of r from X and Y; deviations from the mean.

Country	X	Y	X^2	Y^2	XY
Denmark............	7.2	17.7	51.84	313.29	127.44
France..............	11.7	21.1	136.89	445.21	246.87
Holland.............	8.7	19 2	75.69	368.64	167.04
Italy...............	15.2	22.5	231.04	506 25	342.00
Norway.............	3.8	17.8	14.44	316.84	67.64
Sweden.............	5.1	18.0	26.01	324 00	91.80
Switzerland..........	8.6	19.7	73.96	388.09	169.42
Total.............	60.3	136.0	609.87	2662.32	1212.21
AM.............. =	8.614	19.429			
AM^2............. =	74.2010	377.4860			

$$r = (\Sigma XY - nM_xM_y) \div [(\Sigma X^2 - nM_x^2)(\Sigma Y^2 - nM_y^2)]^{1/2}$$
$$= (1212.21 - 7 \times 8.614 \times 19.429) \div [(609.87 - 7 \times 74.201)(2662.32 - 7 \times 377.486)]^{1/2}$$
$$= 40.68/(90.46 \times 19.92)^{1/2} = 40.68 \div 42.45 = 0.96.$$

Correlation of tabulated data (*Sm*).—When extensive data are used in correlation problems, it is necessary to tabulate them in a double frequency table such as was suggested in Chapter II, p. 21. The correlation is worked out from such a table on the same principles as before, but the arithmetic process must be adapted to the new form in which the data appear.*

The calculation of the coefficient of similarity is illustrated in Example 69. The correlative data are first tabulated as indicated in I. That is, the X-series is tabulated in classes having the limits 1 to 3; 3 to 5; 5 to 7; 7 to 9; and 9 to 11; the Y-series is tabulated in classes having the limits 2.5 to 7.5; 7.5 to 12.5; 12.5 to 17.5; 17.5 to 22.5; and 22.5 to 27.5. The tabulation in the double frequency form combines both of the separate tabulations, the X-frequencies appearing as the footings of the columns and the Y-frequencies appearing at the right as the footings of the rows. It will be noted that the Y-tabulation is

* The simplest form of the correlation table, or scatter diagram, is one in which each set of data is divided into two classes. For example, suppose that 100 cities are classified as large or small in respect to population, and rich or poor in respect to per capita wealth, thus:

	Small	Large
Rich...........	(a) 10	(b) 30
Poor...........	(c) 35	(d) 25

In such a case the coefficient (ω) may be estimated as:

$$\omega = \frac{\sqrt{bc} - \sqrt{ad}}{\sqrt{bc} + \sqrt{ad}} \quad \frac{32.4 - 15.8}{32.4 + 15.8} = 0.344$$

Such a method of correlation has the advantage that it can be applied to qualitative data. But it is, of course, merely a crude method of approximation.

inverted in order to bring the largest class mark in the top row, in accordance with a graph of the correlation. If the data were time series requiring trends, the table would be in terms of deviations from the respective trends. In either case the class marks may be written in unit intervals, if this is more convenient, without changing the final result.

The next step after tabulating the data is the calculation of the average and the average deviation for each tabulation separately. This step appears in Part II of Example 69, where the short-cut method is employed. One step is added to the usual calculation, namely, the finding of the deviations ($d = m - AM$) and the changing of these deviations to units of the average deviation (d/AD) by dividing by the appropriate average deviation. The two scales of deviations thus obtained are used in the final tabulation from which the coefficient is calculated.

Part III of Example 69 illustrates the method of computing the coefficient of similarity. The procedure appears a little complex at first, but in reality is comparatively simple. The frequencies in the tabulation may be copied directly from the primary tabulation in Part I as they represent the same distribution with a mere change of scale. It is now necessary to summate the numerically smaller of each of the correlatives, taking account of the proper sign of correlation. This is most easily accomplished as follows: Following each frequency, write x or y to indicate which of the two correlatives (d/AD; x- or y-scale) is numerically the smaller, and prefix to the frequency the sign of the numerically *larger* correlative. For example, the frequency 1, appearing in the lower left-hand corner, is labeled y because 2.07 in the y-scale is numerically smaller than 2.50 in the x-scale, and the minus sign is prefixed because the larger correlative (-2.50) is negative. In a similar way each frequency is given a sign, either plus or minus, and is labeled either x or y. The x's are totaled by columns and the y's by rows to give the footings Σf_x and Σf_y, respectively. The footing of each column is then multiplied by the d/AD at the head of the column, and the footing of each row is multiplied by the d/AD in the y-scale at the left. The resulting products represent partial sums of the smaller correlatives (Σs) and when totaled give $\Sigma s = 10.78 + 2.50 = 13.28$. This total divided by the number of cities ($n = 20$) is the coefficient of similarity ($Sm = \Sigma s/n = 0.664$). The corresponding Pearsonian coefficient of correlation, computed on the basis of this coefficient of similarity, is $r = 0.829$.

Example 69.—Computation of the coefficient of similarity (Sm), from data assumed to measure certain correlative conditions (X and Y) in cities as indicated. (I) The data are entered in a double frequency tabulation having X classes 1–3;

3–5; etc., and Y classes 2.5–7.5; 7.5–12.5; etc. (II) The means (AM) and average deviations (AD) of the X and Y tabulations are found, and the class marks are expressed as deviations from the mean, reduced to units of the respective average deviations (d/AD). (III) The double frequency table is rewritten with the X- and Y-scales reduced to d/AD units. Each frequency in this tabulation is labeled x or y to indicate the numerically smaller correlative d/AD, and the sign of the larger d/AD is prefixed. The x-frequencies are next totaled by columns and the y-frequencies by rows to give Σf_x and Σf_y. Each of these partial totals is next multiplied by its d/AD. The grand total of the results thus obtained represents Σs, from which the coefficient of similarity ($Sm = \Sigma s/n = 0.664$) is then obtained. The corresponding value of r is found by the formula $r = (2Sm^2 - Sm^4)^{1/2} = 0.829$.

I. Data and tabulation.

City	X	Y	City	X	Y	City	X	Y	City	X	Y
A...	6 7	13	*F*....	4 0	5	*K*...	5 4	22	*P*....	8.8	13
B...	8.0	20	*G* ...	2 0	5	*L*....	4.0	15	*Q*....	3.5	8
C....	4 2	11	*H*...	5 6	9	*M*...	6.6	18	*R*....	6.4	11
D...	10.0	25	*I*....	7.2	17	*N*...	8.0	25	*S*....	4.5	12
E...	6.0	17	*J*....	10 0	20	*O*....	5.3	15	*T*....	3.8	9

Tabulation of X and Y in double frequency table

X class marks

Y class marks	2	4	6	8	10	f
25	..	..	..	1	1	2
20	..	..	2	1	1	4
15	..	1	3	2	..	6
10	..	4	2	..	..	6
5	1	1	..	..	..	2
f	1	6	7	4	2	20

II. Means and average deviations.

X-Tabulation

m	f	mf		d	d/AD
2	1	2		−4	−2.50
4 AM=(6.0)	6(7) bal=(6)	24 (36)	 62	−2	−1.25
6.	7(13)	42		0	0
8.	4	32		2	1.25
10.	2	20	94	4	2.50
	n=20	Σ=120	Σ′d′=32		

AM=6 0 AD=1.6

Y-Tabulation

m	f	mf		d	d/AD
5	2	10		−9.5	−2.07
10	6(8)	60		−4.5	−0.98
AM = (14.5)	bal = (4)	(58)	128		
15	6(12)	90		0.5	0.11
20	4	80		5.5	1.20
25	2	50	220	10.5	2.28
	$n = 20$	$\Sigma = 290$	$\Sigma' d' = 92$		

$AM = 14.5 \qquad AD = 4.6$

III. Coefficient of similarity (negative signs of frequencies are merely indicators for Σs).

x-deviations, AD units

y-deviations, AD units	d/AD	−2.50	−1.25	0	1.25	2.50	Σf_y	$\Sigma f_y y$
	2.28	......	......	....	$1x$	$1y$	$1y$	2.28
	1.20	......	......	$2x$	$1y$	$1y$	$2y$	2.40
	0.11	......	$-1y$	$3x$	$2y$	....	$1y$	0.11
	−0.98	......	$-4y$	$-2x$	....	....	$-4y$	3.92
	−2.07	$-1y$	$-1x$	....	....	....	$-1y$	2.07
	$\Sigma f_x =$	0	$1x$	$3x$	$1x$	0	(Total)	10 78
	$\Sigma f_x x =$	0	1.25	0	1.25	0	2.50	20)13.28
							$Sm =$	0.664

Correlation of tabulated data (r).—The calculation of the Pearsonian coefficient of correlation from tabulated data is analogous to the method already explained for untabulated data (Example 68, p. 240), but the computations are applied to a double frequency distribution of the data if the normal is a mean, or of the deviations from a trended normal. The process is illustrated in Example 70, which is based on the same data as were used in the preceding discussion of the coefficient of similarity (Example 69, p. 242). The double frequency tabulation is prepared from the data as before, and the standard deviations of each of the two distributions, x and y, are computed. These calculations are shown in connection with the double frequency table, although it might be more convenient, as far as the form is concerned, to write the x-distribution in a separate table. The deviations are taken from an assumed average in terms of the actual class intervals, but it is usually

more satisfactory to express the deviations in units of class intervals (i.e., $x = -2, -1, 0, +1, +2$). Such a change of scale will not affect the coefficient of correlation, hence no correction need be made for it unless the averages and standard deviations are required in the original units. In this case the c and σ of each distribution are multiplied by the appropriate class interval (i).

When the measures of dispersion have been determined, the expression Σxy may be obtained as follows: In the first row of the tabulation, multiply each frequency (f_r) in that row by its corresponding x-value, and enter the total under the column $\Sigma f_r x$ (e.g., $1 \times 2 + 1 \times 4 = 6$); similarly find the expression $\Sigma f_r x$ for each row. Then multiply each item in the column thus obtained ($\Sigma f_r x$) by the y-value for that row as written in the left-hand margin of the table. The products are obviously $\Sigma f_r xy$ for each row, and the footing of the column is Σxy, as measured from assumed origins in each scale.

In applying the formula for the coefficient of correlation (r), a correction must be made for the fact that the deviations x and y are from assumed origins rather than from the respective means. The correction follows the logic of the second formula used in Example 68, II (p. 241), and subtracts from Σxy the expression nc_xc_y, the corrections c_x and c_y being the means of the x and y distributions, respectively. It will be found more convenient to divide the n which appears in the denominator of the formula for r into the numerator thus corrected, giving the formula

$$r = (\Sigma xy/n - c_xc_y)/(\sigma_x\sigma_y)$$

If, however, the measures of dispersion σ_x and σ_y are not of interest, the formula may be further abbreviated to the form

$$r = (\Sigma xy - nc_xc_y)/[(\Sigma x^2 - nc_x{}^2)(\Sigma y^2 - nc_y{}^2)]^{1/2}$$

It will be seen that, as far as the numerator of the formula is concerned, the correction c_x times c_y may be disregarded if either one of the factors is zero, or is so small that the product becomes negligible. Hence if one of the distributions x or y is taken from the mean, the correction of the numerator (nc_xc_y) may be disregarded.

Example 70.—Computation of Pearsonian coefficient of correlation (r) from tabulated data. The double frequency table is made up from the data of Example 69 (p. 242). Scales x and y are written as deviations from assumed means (they may be written as step deviations: 1, 2, 3, 4, 5, or $-2, -1, 0, +1, +2$, without affecting r). The standard deviation of each distribution is calculated by the usual short-cut method, and for the y-series $\Sigma f_r x$ is obtained for each row (f_r is each partial frequency in the row). Multiplying the column $\Sigma f_r x$ by y gives $\Sigma f_r xy$, the total of which is Σxy, as in untabulated data. Since x and y are not centered deviations, Σxy must be corrected by subtracting nc_xc_y as in Example 68, II, p. 241 (c_x and c_y are the

means of x and y). The formula $(\Sigma xy - nc_xc_y) \div [(\Sigma x^2 - nc_x^2)(\Sigma y^2 - nc_y^2)]^{1/2}$ is more conveniently written as first given below.

	$X =$	2	4	6 (A_a)	8	10					
Y	y	$x = -4$	-2	0	2	4	f	fy	fy^2	$\Sigma f_r x$	$\Sigma f_r xy$
25	10	...	....	...	1	1	2	20	200	6	60
20	5	...	....	2	1	1	4	20	100	6	30
$A_a = 15$	0	...	1	3	2	....	6	0	0	2	0
10	-5	...	4	2	....	....	6	-30	150	-8	40
5	-10	1	1	...	....	....	2	-20	200	-6	60
	f	1	$+6$	$+7$	$+4$	$+2$	$=20$	$)-10$	650		$20)190 = \Sigma xy$
	fx	-4	-12	0	$+8$	$+8$	$= 0$	$-0.5 = c_y$	32.5		$9.5 = \Sigma xy/n$
							$c_x = 0$				
	fx^2	16	$+24$	$+0$	$+16$	$+32$	$= 88 = \Sigma x^2$		$c_y^2 = .25$		
							$\sigma_x^2 = 4.4$		32.25		
							$\sigma_x = 2.098$		$\sigma_y = 5.679$		

$$r = (\Sigma xy/n - c_xc_y) \div (\sigma_x\sigma_y) = (9.5 - 0) \div (2.098 \times 5.679)$$
$$= 9.5/11.9145 = 0.797.$$

Or
$$r = (\Sigma xy - nc_xc_y) \div [(\Sigma x^2 - nc_x^2)(\Sigma y^2 - nc_y^2)]^{1/2}$$
$$= (190 + 20 \times 0 \times 0.5) \div [(88 - 20 \times 0)(650 - 20 \times 0.25)]^{1/2}$$
$$= 190 \div (88 \times 645)^{1/2} = 190 \div 238.24 = 0.798.$$

Correlation by diagonal deviations.*—The coefficient r may be computed by means of the diagonal deviations, σ_v and σ_w, previously mentioned. If the double frequency table of Example 70 is scaled to unit intervals ($x = 1, 2, 3$, etc.), four frequency tabulations may be written: (1) down the columns gives the x-frequencies, or 1, 6, 7, 4, 2; (2) across the rows gives the y-frequencies, or 2, 4, 6, 6, 2; (3) diagonally downward from upper left gives the v-frequencies, or 1, 1, 4, 3, 3, 4, 1, 2, 1; and (4) diagonally downward from upper right gives the w-frequencies, or 4, 10, 6. The variances (σ^2) calculated from these frequencies taken as representing successive classes, where $i = 1$, are $\sigma_x^2 = 1.1$; $\sigma_y^2 = 1.29$; $\sigma_v^2 = 4.29$; and $\sigma_w^2 = 0.49$. It may readily be shown that $\frac{1}{4}(\sigma_v^2 - \sigma_w^2) = \Sigma xy/n$, taken in a unit class frame, as above.

Hence (unit class intervals),

$$r = (\sigma_v^2 - \sigma_w^2) \div (4\sigma_x\sigma_y) = (4.29 - 0.49) \div (4\sqrt{1.1 \times 1.29})$$
$$= 3.80 \div (4 \times 1.19122) = 0.7975.$$

Since $2(\sigma_x^2 + \sigma_y^2) = (\sigma_v^2 + \sigma_w^2)$ the formula may be written:

$$r = (\sigma_x^2 + \sigma_y^2 - \sigma_w^2) \div (2\sigma_x\sigma_y)$$

in which form it is convenient for calculation. The expression $\sigma_x\sigma_y$ is

* The principle here described has been elaborated into a convenient method of calculation of the coefficient of correlation, r, by E. E. Cureton and J. W. Dunlap. A chart arranged for computations, entitled "C–D Machine Correlation Chart," is published by the Macmillan Company and will be found very useful in laboratory work.

most readily obtainable as the geometric mean of the x and y variances, that is, $\sqrt{\sigma_x^2\sigma_y^2}$. In negative correlations the work may often be abbreviated by substituting σ_v^2 for σ_w^2, but when this is done the sign of correlation is reversed.

The method of rank-differences.—The Pearson method is sometimes applied to the rankings of the data, a procedure which is analogous to using the median as the type, and the quartile deviation as the measure of dispersion. It is, therefore, especially suitable for use with irregular data where extreme items are erratic and perhaps inaccurate. It is applied to untabulated rather than tabulated data.*

The ranking method implies substituting for the data their rank, as highest (rank 1); next highest (rank 2), etc. If this order of ranking is reversed in one series, the sign of correlation is reversed—a time-saving device with presumptively negative correlations. If ties occur which cannot be resolved by a further study of the data, they may be assigned the average of the ranks they would have received if they had differed slightly. Thus, if two items are tied for second place, each would be ranked $(2 + 3)/2$ or 2.5, and the next item in order of size would be ranked fourth. This adjustment for ties is perhaps as satisfactory as any, but it is not entirely valid; and if the ties are numerous, the result is less reliable. The process is illustrated in Example 71, and the rankings and the regression line are plotted in Chart 37. In time series, the deviations from normal may similarly be ranked from the largest positive down to the largest negative. The formula, $r_r = 1 - 6\Sigma d^2 \div [n(n^2 - 1)]$, is an algebraic simplification of the usual formulas for the Pearsonian coefficient as applied to rankings of the data.

Example 71.—Correlation by the method of ranks. $r_r = 1 - 6\Sigma d^2 \div [n(n^2 - 1)]$, where d is the difference in ranks of each pair of correlatives. The data are ranked from the largest to the smallest in each series respectively, and the differences (d) in ranks taken. The formula applied to these differences is an algebraic simplification of the formula $r = \Sigma xy \div (n\sigma_x\sigma_y)$ as applied to rankings. If ties occur in rankings

* A still simpler but rather crude method of correlation, which is sometimes useful in a preliminary investigation, is the method of concurrent deviations. Each deviation in the two correlative series is first labeled plus or minus, with respect to the preceding item, the two adjacent items, the normal, or any other desired basis. These plus and minus series are then correlated by counting the number of agreements (+ with +, or − with −) and the number of disagreements (+ with −, or − with +). If the former number is larger, the correlation is positive; if smaller, it is negative. In either case the larger sum is labeled C (number of concurrences), and the coefficient is found as follows:

$$\text{Coef.} = \pm [(2C - n) \div n]^{\frac{1}{2}}$$

where n is the number of pairs of correlative items.

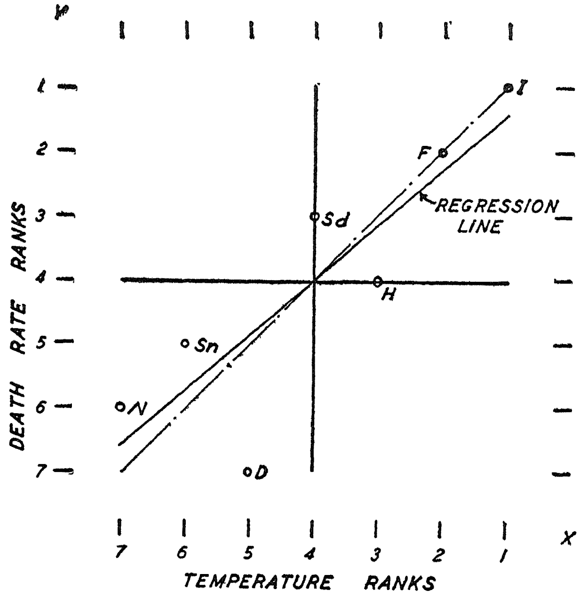


CHART 37

Scatter diagram of rankings of certain countries in death rates (Y) and mean annual temperature (X). For the data see Example 71. The coefficient of correlation is $r = 0.86$, which is the slope of the regression line. This line is in effect a straight-line trend fitted to the data, as plotted, by the method of least squares. The standard deviation (S) of the plotted data from this regression line is $S = [\Sigma(Y - T)^2/n]^{1/2} = 1.03$. This measure is known as the standard error of estimate, and in linear correlation is related to the coefficient of correlation by the equation $r^2 = 1 - S^2/\sigma_y^2$.

they are first written consecutively and then averaged. But to the extent that ties occur the result is less valid. The abbreviated data represent mean annual temperature (X) in degrees Centigrade and annual death rates (Y) per 1000 population.

	Data		Ranks			
	X	Y	X	Y	d	d^2
Denmark............	7.2	17.7	5	7	−2	4
France..............	11.7	21.1	2	2	0	0
Holland.............	8.7	19.2	3	4	−1	1
Italy...............	15.2	22.5	1	1	0	0
Norway..............	3.8	17.8	7	6	1	1
Sweden..............	5.1	18.0	6	5	1	1
Switzerland.........	8.6	19.7	4	3	1	1
					0	8

$$r_r = 1 - \overline{6\Sigma d^2 \div [n(n^2 - 1)]} = 1 - 6 \times 8 \div (7 \times 48) = 1 - 0.14 = 0.86.$$

The regression line.—In Chart 37 the degree of correlation is expressed by a regression line having the slope of $r = 0.86$. It may

be proved that when the x- and y-scales of correlative data have the same standard deviation, as in the case of rankings, or when each scale has been reduced to units of its own standard deviation, the Pearsonian coefficient of correlation is simply the slope of the straight-line trend fitted to the data by the method of least squares. This trend is easily located in the chart, after drawing the axes at the X and Y averages, by measuring to the right from the intersection of the axes any convenient distance taken as a unit, and then measuring vertically the fraction of this unit represented by r. The point thus determined will lie on the trend (regression line), which may then be drawn through this point and through the intersection of the axes.

If the X and Y data do not have the same spread, that is, if σ_x is not equal to σ_y, the slope (b) of the regression line or trend may be easily computed from r by making allowance for the two standard deviations, σ_x and σ_y, as follows:

$$b = r\sigma_y/\sigma_x$$

and in a like manner the value of r may be found if a line of least squares has been fitted to the data, as follows:

$$r = b\sigma_x/\sigma_y$$

It may be seen from these formulas that r and b are necessarily identical when σ_x and σ_y are identical.

A relation will also be found to exist between the deviations of the y data from the regression line, or trend as thus computed. It is readily seen that if the correlation is $+1.00$ the data as plotted will lie exactly upon the trend, and the deviations consequently will be 0. But as the deviations increase, the degree of correlation diminishes. The relation of the deviations to r may be expressed by finding the standard deviation of the points as plotted from the regression line $(\Sigma\overline{Y - T}^2/n)^{1/2}$. This standard deviation is known as the standard error of estimate (S), and $0.6745S$ is the probable error of estimate. The term "error of estimate" was originally suggested by the fact that correlations are sometimes used to estimate the dependent variable (Y) when the independent variable (X) is known, in which case the standard error of estimate (S) suggests the relative accuracy of such an estimate. The relation of r to S, where S is measured on the Y-ordinates, may be proved mathematically to be (square roots are plus or minus):

$$r = (1 - S^2/\sigma_y^2)^{1/2}$$

It is sometimes desirable to express the regression line as a trend in terms of the original X and Y data, rather than in terms of the axes

and the slope. This may readily be done by writing the regression line, or trend (line of least squares), as

$$T = a + bx$$

where x has its origin at AM_x, and a is AM_y at this origin. Since $x = X - AM_x$, this equation becomes

$$\begin{aligned} T &= AM_y + b(X - AM_x) \\ &= AM_y - bAM_x + bX \end{aligned}$$

When the data are very complete and the correlation is likely to be typical of the general relationship existing between X and Y, T (the (regression line) as derived from the above equation may be taken to indicate an estimate of Y for any correlative magnitude of X. If X and Y are mutually dependent, the X may be similarly estimated from Y, by drawing another regression line derived on the assumption that Y is the independent series. In a scatter diagram of the standard deviation units, this second regression line has the same slope as the first, but on a different base; hence the two regression lines do not coincide except when each slope is unity. However, as has already been indicated, such estimates are significant only when r has a high value, and when the correlation as measured may be assumed to have a general applicability. Such conditions are not commonly met with in social statistics, and as a result correlations should generally be taken merely as measurements of specific interrelationships, rather than of general laws.

Curvilinear correlation.—The discussion of correlation thus far has been based on the assumption that the regression is linear; that is, that a straight-line trend fitted to the Y-data constitutes the most appropriate regression curve. This means that as one of the variables increases the other tends either to increase or decrease at a constant rate throughout the limits of the data.

It will readily be seen that if the data are carried to extremes, linear regression is likely to be materially modified. For example, within certain limits an increase of rainfall will increase crop yields; but after a certain optimum point has been reached, additional rainfall is likely to cause a decline in yields. Similar conditions are likely to obtain in many social and economic situations. Hence when correlation covers rather wide limits it must often be measured by a curved rather than a straight line. Such a process of measurement is known as curvilinear correlation.

The process of computing a coefficient of curvilinear correlation is comparatively simple in principle but often becomes very complex in practice. In the first place it must be recognized that curvilinear correlation is presumably a measurement of a cause and effect relationship, or of series that are indirectly connected by such a relationship. The series which is regarded as the causal factor, or which is a function of that factor, is taken as X, and the series which is regarded as an effect, or as a function related to a given effect, is taken as Y. The correlation is assumed to express the regression of Y on X. For example, in Chart 38, p. 258, there are plotted the supposed crop yields in a number of comparable farms under conditions of varying rainfall. It is assumed that the regression of crop yields on rainfall is fairly represented by a parabola with a negative curvature, that is, by a line which rises as rainfall increases from a minimum, but which decreases after a certain optimum rainfall is reached. Although the hypothetical data do not seem to bear out this assumption very exactly, yet the parabolic trend fitted to the data is taken to represent the general tendency in this particular study.

Before the fitting of a parabolic regression curve is studied, a simple case in which the regression is expressed by the averages of the columns in tabulated data may be considered. A coefficient of correlation obtained on the basis of such a regression is usually called a correlation ratio. It is not, however, a very satisfactory measure, partly because it depends too largely upon the grouping of the data in columns and partly because it tends to follow the random fluctuations of the data rather too closely and so may fail to register the general trend. It therefore has a tendency to exaggerate the degree of correlation. This tendency may, however, be minimized if care is taken in arranging the columns so that column averages form a fairly constant curve.

At the beginning of this chapter it was seen that correlation is essentially measured by the deviations of the data from the diagonals in the squared scatter diagram. That is, linear correlation is a function of the degree of scatter of the data from such a straight line. It has also been seen that there is a mathematical relation between the standard error of estimate (S), which measures the deviations of the data from the trend of linear regression, and the Pearsonian coefficient of correlation (r). This relationship is expressed by the formula

$$r = (1 - S^2/\sigma_y^2)^{1/2}$$

where S is measured on the Y-ordinates. It will be seen, however, that

if r is measured by this formula the sign of correlation is not determined, though it may be readily determined by reference to the slope of the regression line.

The measurement of correlation by the use of the formula expressing the relationship of r and S, as just given, may be applied to curvilinear as well as to linear correlation. Such a coefficient (ρ) is the general measure of correlation of which the Pearsonian r is merely a specific case. As just noted, the result will be an index ranging from 0 to +1, measuring the closeness of fit of the regression curve to the data. A coefficient of 1 means that the data fall exactly upon the regression curve; a coefficient of 0 means that no regression curve other than a horizontal trend is appropriate; that is, the data show no definite trend. In curvilinear correlation the sign is taken as positive, since in such a case a negative coefficient obviously has no special significance.

There are certain other aspects of general correlation which should be noted. In the first place, it may be proved that for most trends, including those formed by averages of columns, and those fitted as least squares parabolas, the value of ρ is identical with the value of r obtained by the linear correlation of the data and the regression curve. Such a correlation, it will be seen, is a measure of the closeness of fit of the trend. Likewise from another point of view the general correlation may be considered as a comparison between the scatter of the data (σ_y) and the scatter of the trend (σ_t); that is, $\rho = \sigma_t/\sigma_y$. This relationship is true for all cases in which the previous formula holds (cases where $\Sigma YT = \Sigma T^2$). These two modifications of the basic correlation formula will be found useful in the illustrations that follow. Even where they are not precise equivalents of ρ, they may be taken as convenient approximations.

The coefficient of similarity as a correlation ratio (η_s).—In applying the coefficient of similarity to curvilinear regression the most convenient and appropriate method consists of the linear correlation of the data with the regression curve. This procedure is illustrated in Example 72, where the regression curve is taken as the means of the columns. That is, the result is a form of the correlation ratio. The data chosen for this illustration do not, however, show enough curvature to be significant. Hence the result obtained is only a trifle larger than that previously obtained by linear correlation. As a rule, however, the method will give a larger result and, as previously noted, sometimes to a misleading degree.

The process of computation is probably sufficiently explained in the example itself. The x- and y-scales are expressed in units of their

respective average deviations as already found (cf. Example 69, p. 242). The means of the columns are calculated by taking an average of the y-scale weighted with the frequencies appearing in the given column. Thus in the first column the average is merely -2.07 because the only frequency in the column appears at that point on the scale. The average of the second column is $1(0.11) + 4(-0.98) + 1(-2.07)$ divided by the sum of the frequencies or 6, which gives -0.98. This result is, of course, obvious from the symmetrical arrangement of the columns. The other columns are similarly averaged and the results entered in the line below the footings of the frequencies. It will be seen that the negative signs appearing before some of the frequencies have no significance for this part of the operation, since they are merely indicators used later in determining the sign of the correlation.

The average deviation of the column means (M_c) is next found, the column frequencies (f) being taken as weights; and the means themselves are then divided by their average deviation ($M_c \div AD$). The scale thus obtained is now taken as the x-scale instead of that appearing at the top of the table. It is, of course, the regression curve expressed in units of its own average deviation, and the calculation that follows is the correlation of the y-data with this regression curve. The designation of each frequency as x or y to indicate the smaller correlative, and the prefixing of a negative sign when the larger correlative has that sign, follow the same plan used before in linear correlation (cf. Example 69, p. 242). The totaling of the y-deviations by rows and their multiplication by the appropriate y, together with a similar totaling of the x-frequencies by columns and their multiplication by the appropriate x, give products which total as Σs, and this, divided by n, gives the curvilinear coefficient of similarity.

Example 72.—Coefficient of similarity (η_s), curvilinear correlation employing the averages (M_c) of the y-columns as the regression (trend) line. The y-average (M_c) of each column is found by weighting the y's with the column frequencies (positive). The average deviation of the M_c distribution, weighted by f, is obtained as indicated ($\Sigma' f M_c'/n$). The M_c distribution is reduced to units of this average deviation, and as thus reduced (M_c/AD) is taken as x in the computation that follows. Next, the frequencies in the correlation table are labeled x or y indicating which of the correlatives is numerically the smaller, and the sign of the numerically larger correlative is prefixed. The frequencies labeled y are totaled by rows to the right (f_y), and the frequencies labeled x are totaled by columns (f_x), and $f_y y$ and $f_x x$ are computed and totaled. The grand total is Σs, which, divided by $n = 20$, gives η_s, the curvilinear coefficient of similarity.

Double frequency table, AD units; see Example 69, Part III, p. 244
(Negative signs of frequencies are merely indicators for Σs)

Original x-deviations, AD units
M_c/AD is taken as x in the correlation

Y-deviations, AD units	d/AD	-2.50	-1.25	0	1.25	2.50	f_y	$f_y y$
	2.28	......	......	......	$1x$	$1x$	0	0
	1.20	......	......	$2x$	$1x$	$1y$	1	1.20
	0.11	......	$-1y$	$3y$	$2y$	....	4	0.44
	-0.98	......	$-4y$	$-2x$	....	....	-4	3.92
	-2.07	$-1y$	$-1x$	......	....	....	-1	2.07
	f	1	6	7	4	2		7.63
	M_c	-2.07	-0.98	0.11	0.92	1.74		
	fM_c	-2.07	-5.88	0.77	3.68	3.48	15.88	
						$AD=$	0.794	
	$M_c/AD=x$	-2.61	-1.23	0.14	1.16	2.19		
	f_x	0	-1	0	2	1		
	$f_x x$	0	1.23	0	2.32	2.19		5.74

$20\overline{)13.37} = \Sigma s$

$\eta_s = 0.6685$

$r = 0.83 = \eta$

$(r^2 = 2Sm^2 - Sm^4)$

The correlation ratio (η).—The value of ρ computed by an analogous method, on the basis of a like regression curve, is illustrated in Example 73. In this case it is more convenient to use the relationship $\rho = \sigma_t/\sigma_y$ (the ratio of the standard deviation of the trend to the standard deviation of the Y's, each measured on Y ordinates). The value of σ_y is obtained from the double frequency distribution in the usual manner by assuming an average and making corrections in Σy^2, as indicated, and σ_t may be found by first averaging the columns in a manner similar to that just explained. This average may be made on the basis of the original Y-scale or of the deviation y-scale as desired. The standard deviation of these means weighted with the respective frequencies of the column is calculated. This is found to be 4.55 as compared with a y-standard deviation of 5.68. The ratio of these two standard deviations is 0.80, which is ρ as computed from the regression curve of column averages, or the correlation ratio (η) as it is commonly known. The example in Part II illustrates an alternate form for finding σ_t which will often prove to be a little shorter. The limitations upon this method of curvilinear correlation have already been discussed.

Example 73.—Computation of the correlation ratio (η), by a formula of non-linear correlation. In effect a trend is fitted to the data of the regression table, and the standard deviation (σ_t) of this trend about the Y-mean (which is the same as the trend mean) is compared with the standard deviation of the Y-distribution; that is, the correlation ratio is $\eta = \sigma_t/\sigma_y$. The trend employed consists of the means (M_c) of the columns. The mean (M_c) of any column is $\Sigma f_c Y/f$ for that column (f_c being the frequencies in the column, and f the footing), and obviously $\Sigma f_c Y = fM_c$. Hence the work may be abbreviated by writing fM_c as $\Sigma f_c Y$, and taking fM_c^2 as $(\Sigma f_c Y)^2/f$, thus eliminating the calculation of M_c. The calculation is usually briefer, also, if based on deviations (y) from an assumed mean instead of Y, as in Part II.

Tabulation for computation of $\eta = \sigma_t/\sigma_y$ (for data see Example 69, p. 242)

Column No.			1	2	3	4	5			
		X	2	4	6	8	10			
					M_a					
		x	−4	−2	0	2	4	f	fy	fy^2
Y	y									
25	10	. .	. . .	. . .	. .	1	1	2	20	200
20	5	. .	. . .	. . .	2	1	1	4	20	100
M_a = 15	0	. .	. . .	1	3	2	. .	6	0	0
10	− 5	. .	. . .	4	2	. .	. .	6	−30	150
5	−10	. .	1	1	. .	. .	. .	2	−20	200
		f	1	6	7	4	2	20	)−10	650
									c = −0.5	32.5
										c^2 = 0.25
										σ_y^2 = 32.25
										σ_y = 5.68

I. σ_t; or Y-dispersion about column means (M)

Column	M_c	f	fM_c	fM_c^2
1	5	1	5	25
2	10	6	60	600
3	15	7	105	1575
4	18.75	4	75	1406.25
5	22.50	2	45	1012.50
		20	)290	4618.75
		AM =	14.5	230.94
			AM^2 =	210.25
			σ_t^2 =	20.69
			σ_t =	4.55

II. Alternate form, substituting y (deviations from assumed mean) for Y, and not finding M

Column	f	$\Sigma f_c y$	$(\Sigma f_c y)^2/f$
1	1	−10	100
2	6	−30	150
3	7	0	0
4	4	15	56.25
5	2	15	112.50
	20	)−10	418 75
		c = −0.5	20.94
		c^2 =	.25
		σ_t^2 =	20.69
		σ_t =	4.55

Correlation ratio (η):

$$\eta = \sigma_t/\sigma_y = 4.55/5.68 = 0.801$$

Curvilinear correlation; parabolic regression.—Reference has already been made to the case of curvilinear correlation furnished by data of crop yields as a regression on rainfall, and it was noted that a quadratic parabola might form a suitable regression curve in this case. The computation of the regression equation in such a problem will be the same whether the coefficient is to be calculated on the basis of average deviations or standard deviations; that is, whether Sm_c (cf. p. 258) or ρ is to be found. The calculation of the equation of the parabolic trend is illustrated in Example 74. The most exact method of such a calculation would require untabulated data expressing the yield and the rainfall during the growing season for each farm separately. But it is usually sufficient to calculate the trend on the basis of the averages of suitably tabulated columns, rather than from the specific items. In either case, however, the computation of the trend requires a somewhat longer process than that used in ordinary time trends where the regularity of the data allow the use of short-cut formulas. The method must therefore be developed on the basis of the normal equations (cf. p. 165). It is possible to calculate from the data all the elements of the normal equations with the exception of the constants, a, b, and c, and these constants may be obtained by an algebraic solution of the equations. In the solution of the equation on the basis of column averages, it is, of course, necessary to use the column frequencies as weights. By substituting the required elements in the normal equations, the values of a, b, and c may be found by any convenient algebraic solution of simultaneous equations. After the constants have been determined, the parabolic trend may be computed as indicated. The trend points thus found have, of course, the weights of the appropriate frequencies. On the basis of the regression curve thus determined, the degree of correlation may be measured by either Sm_c or ρ.

Example 74.—Curvilinear regression. Assumed data of inches of rainfall (X) in growing season and resulting bushels per acre (Y) for 32 representative farms in a district having varying degrees of rainfall. A parabolic trend is fitted to the column averages weighted with the column frequencies. This is done by means of the normal equations of the parabolic trend, since the formulas for the constants previously used do not take account of frequencies. All summations must employ the frequencies. The constants are obtained by substituting the required functions of x and Y in the normal equations and solving for a, b, and c. The trend is then computed as $T = a + bx + cx^2$ for the given magnitudes of x. The trend thus found is taken as a regression curve representing the probable number of bushels per acre for given inches of rainfall (x) within the area studied.

Inches of rainfall in growing season

Bushels per acre (Y)	(X) 2	6	10	14	18	f
25	..	..	..	1	..	1
24	..	..	..	2	..	2
23	..	..	..	2	..	2
22	..	..	..	2	..	2
21	..	..	1	1	..	2
20	..	..	1	..	..	1
19	..	1	2	..	..	3
18	..	2	4	..	..	6
17	..	2	2	..	..	4
16	..	2	1	..	..	3
15	..	1	1	..	..	2
14	..	..	..	..	1	1
13	..	..	..	..	..	0
12	..	..	..	..	..	0
11	..	..	..	..	..	0
10	..	..	..	..	1	1
9	..	..	..	..	..	0
8	..	..	..	..	..	0
7	..	..	..	..	..	0
6	1	..	..	..	..	1
5	..	..	..	..	..	0
4	1	..	..	..	..	1
Totals...	2	8	12	8	2	32

Normal equations:

$$\Sigma Y = na + b\Sigma x + c\Sigma x^2$$

$$\Sigma xY = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$$

$$\Sigma x^2Y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$$

Solution of normal equations (Y is mean of columns):

Y	f	fY	x	fx	x^2	fx^2	x^3	fx^3	x^4	fx^4	xfY	x^2fY
5	2	10	−2	−4	4	8	−8	−16	16	32	− 20	40
17	8	136	−1	−8	1	8	−1	− 8	1	8	−136	136
18	12	216	0	0	0	0	0	0	0	0	0	0
23	8	184	1	8	1	8	1	8	1	8	184	184
12	2	24	2	4	4	8	8	16	16	32	48	96
	32	570	0	0	10	32	0	0	34	80	76	456
		ΣY		Σx		Σx^2		Σx^3		Σx^4	ΣxY	Σx^2Y

Substituting in normal equations:

$$570 = 32a + 0 + 32c$$

$$76 = 0 + 32b + 0$$

$$456 = 32a + 0 + 80c$$

Solution: $b = 2.375$; $c = -2.375$; $a = 20.19$.

Weighted parabolic trend $(a + bx + cx^2)$

x	$T = 20.19 + 2.375x - 2.375x^2 =$			T	f
-2	$= 20.19 - 4.75$	$- 9.5$	$=$	5 94	2
-1	$= 20.19 - 2.375$	$- 2.375$	$=$	15 44	8
0	$= 20.19$		$=$	20.19	12
1	$= 20.19 + 2.375$	$- 2.375$	$=$	20.19	8
2	$= 20.19 + 4.75$	$- 9.5$	$=$	15.44	2

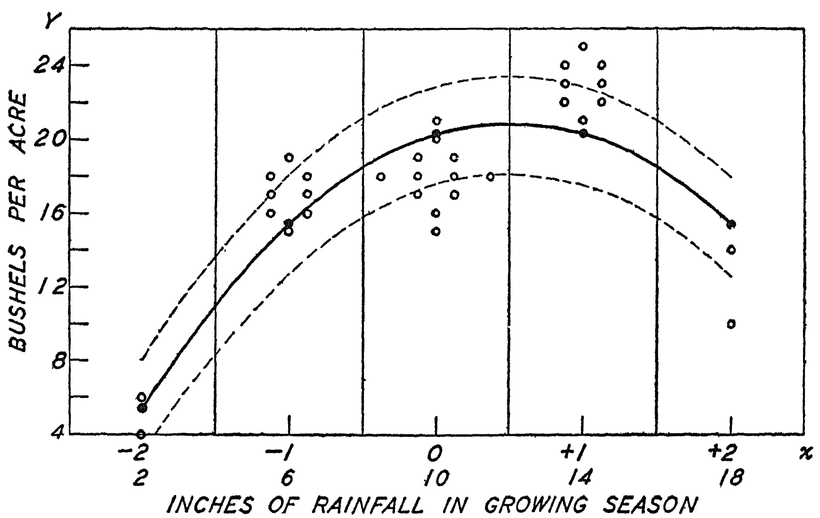


CHART 38

Regression of assumed crop yields on rainfall for 32 comparable farms during a growing season. The data are grouped by columns having class marks of 2, 6, 10, 14, and 18 in. It is assumed that a parabola will here represent with sufficient accuracy the natural regression. The parabola (solid line) is fitted to the column averages weighted by the column frequencies. The calculation of the constants requires the use of the normal equations (cf. Example 74). The solid dots on the regression curve are the points determined by the calculation. The two dotted lines above and below the regression curve represent the standard error of estimate (S), which is the standard deviation of the individual points from the regression curve, measured on the Y-ordinates. The measure of the correlation between rainfall and crop yields is illustrated in Examples 74 and 75.

Coefficient of similarity, curvilinear regression (Sm_c).—The calculation of the curvilinear coefficient of similarity on the basis of a parabolic regression curve is illustrated in Example 75. The process parallels

in almost every particular that illustrated in a previous problem where the regression curve consisted of the column means (cf. Example 72, p. 253). There is, however, some difference in the statement of the problem in that the X- and Y-scales are here written in the original units rather than in average deviation units. There is also the difference that the regression curve (T) here consists of the points on the parabola calculated in the previous problem (cf. Example 74). The process therefore consists, first, in reducing the Y-scale to d/AD units, which is done by the usual procedure. The Y-average is found as $\Sigma Y/n = 570/32 = 17.81$, and the y-deviations are next calculated and divided by their average deviation, $AD = 3.223$. This column is taken as the y-scale for purposes of correlation. The x-scale is next obtained on the basis of the parabolic trend (T) taken as a regression curve. The average of this regression curve weighted with the column frequencies is found as $\Sigma T/n = 570.08/32 = 17.82$. The T-scale is expressed in deviations from this average and is divided by the average deviation. The results thus found are taken as the x-scale. A correlation based on these y- and x-scales will obviously be a correlation of the data and the trend, which is taken as the measure of curvilinear correlation.

The frequencies in the body of the correlation table are next labeled x or y, to indicate which of the correlatives represented by a given frequency is the smaller, and the sign of the larger correlative is prefixed. These signs and the x or y designations are, of course, merely indicators of the correlation and have no significance with respect to the previous calculations of the x- and y-scales. The y-frequencies are totaled (f_y) by rows to the right as before, are then multiplied by y, and totaled to give $\Sigma f_y y$. In the same way $\Sigma f_x x$ is found and added to $\Sigma f_y y$ to give Σs. This total divided by $n = 32$ gives the curvilinear coefficient of similarity, $Sm_c = 0.59$. The corresponding value of r is 0.75, which may be regarded as an estimate of ρ.

Computation of ρ, parabolic regression.—The process of computing directly the value of ρ for the parabolic regression of farm yields on rainfall is illustrated in Example 76. The formula used is the general correlation formula $\rho = (1 - S^2/\sigma_y^2)^{1/2}$, where S is the standard error of estimate calculated on the ordinates. It is possible to use the alternative formula $\rho = r_{yt}$ as in the previous example, or $\rho = \sigma_t/\sigma_y$, either of which formulas would give the same result as the general formula, provided that the parabola had been fitted to each of the Y-points as tabulated for purposes of this correlation. Since, however, the parabola has been fitted to the data by an approximate method, it is safer to use the general formula and probably almost as convenient.

Example 75.—Curvilinear correlation—coefficient of similarity (Sm_c). Regression of crop yields on rainfall (Y = bushels per acre) for farms in comparable areas (X = rainfall in inches during growing season). Regression curve calculated as a parabola trend (T) fitted to weighted column averages (see Example 74). The Y-distribution is reduced to d/AD, which is taken as y; the T-distribution is similarly reduced and taken as x. The frequencies are labeled x or y according to which of the

Double frequency distribution (Negative signs

Y	$X = 2$	6	10	14	18	f	fY
25	....	....	....	$1x$	....	1	25
24	....	....	....	$2x$	....	2	48
23	....	....	....	$2x$	....	2	46
22	....	....	....	$2x$	....	2	44
21	....	....	$1x$	$1x$	....	2	42
20	....	....	$1y$	....	....	1	20
19	....	$-1y$	$2y$	....	....	3	57
18	....	$-2y$	$4y$	....	....	6	108
17	....	$-2y$	$2y$	....	....	4	68
16	....	$-2y$	$1y$	....	....	3	48
15	....	$-1x$	$-1x$	....	....	2	30
14	....	....	....	....	$-1x$	1	14
13	....	....	....	....	....	0	0
12	....	....	....	....	....	0	0
11	....	....	....	....	....	0	0
10	....	....	....	....	$-1x$	1	10
9	....	....	....	....	....	0	0
8	....	....	....	....	....	0	0
7	....	....		....	....	0	0
6	$-1y$	....	....	....	....	1	6
5	....	....	....	....	....	0	0
4	$-1x$	....	....	....	....	1	4
f	2	8	12	8	2	32)570	

$AM_y = 17.81$

T	5.94	15.44	20.19	20.19	15.44	
fT	11.88	123.52	242.28	161.52	30.88	$570.08 = \Sigma fT$
						$17.82 = AM_t$
d_t	−11.88	−2.38	2.37	2.37	−2.38	$94.96 = \Sigma' fd_t'$
						$2.968 = AD$
$d_t/AD = x$	− 4.00	−0.80	0.80	0.80	−0.80	
f_x	− 1	−1	0	8	−2	
$f_x x$	4.00	0.80	0	6.40	1.60	$12.80 = \Sigma f_x x$

correlatives is numerically the smaller; and the sign of the numerically larger is prefixed. The y-frequencies are totaled by rows, and the x-frequencies by columns, and each of these totals is multiplied by its respective y or x. The grand total of these products is Σs, which, divided by $n = 32$, gives Sm_c, the coefficient of similarity. The method parallels that of Example 72, p. 253, except for the calculation of the regression curve.

of frequencies are merely indicators for Σs)

d	$d/AD = y$	f_y	$f_y y$
7.19	2.23		
6.19	1.92		
5.19	1.61		
4.19	1.30		
3.19	0.99		
2.19	0.68	1	0.68
1.19	0.37	1	0.37
0.19	0.06	2	0.12
− 0.81	−0.25	0	
− 1.81	−0.56	−1	1.12
− 2.81	−0.87		
− 3.81	−1.18		
− 4.81	−1.49		
− 5.81	−1.80		
− 6.81	−2.11		
− 7.81	−2.42		
− 8.81	−2.73		
− 9.81	−3.04		
−10.81	−3.35		
−11.81	−3.66	−1	3.66
−12.81	−3.97		
−13.81	−4.29		

$\Sigma' fd' = 103.125$

$AD = 3.223$

$\Sigma f_y y = 5.95$

$\Sigma f_x x = 12.80$

$32\overline{)18.75} = \Sigma s$

$Sm_c = 0.5859$

$r = 0.75 = \rho$

$(r^2 = 2Sm^2 - Sm^4)$

Example 76. **Curvilinear correlation.**—Coefficient of correlation (ρ) for the data of Example 74, p. 256, taking the parabola there computed as the regression curve. Formula $\rho^2 = 1 - S^2/\sigma_y^2$ where S is the standard error of estimate ($S^2 = \Sigma\overline{Y - T}^2/n$). An alternative computation is $\rho = \sigma_t/\sigma_y$, which would give the same result as the preceding if the parabola trend were fitted to the data rather than to weighted averages of the columns. The use of a regression curve fitted to the weighted averages of the Y columns is an approximation which is usually sufficiently exact, and has the advantages of less emphasis on erratic items, and shorter calculations. The calculation of the general coefficient of correlation (ρ) is given below. The regression curve is $T = 20.19 + 2.375\ x - 2.375\ x^2$.

I. Computation of σ_y.

Y	f	$d = Y - 18$	fd	fd^2
25	1	7	7	49
24	2	6	12	72
23	2	5	10	50
22	2	4	8	32
21	2	3	6	18
20	1	2	2	4
19	3	1	3	3
$A_a =$ 18	6	0	0	0
17	4	− 1	− 4	4
16	3	− 2	− 6	12
15	2	− 3	− 6	18
14	1	− 4	− 4	16
13	0	− 5	0	0
12	0	− 6	0	0
11	0	− 7	0	0
10	1	− 8	− 8	64
9	0	− 9	0	0
8	0	−10	0	0
7	0	−11	0	0
6	1	−12	−12	144
5	0	−13	0	0
4	1	−14	−14	196
	32		32)− 6	32)682
			$c = -$ 0.1875	21.3125
				$c^2 =$ 0.0352
				$\sigma_y^2 =$ 21.2773
				$\sigma_y =$ 4.613

II. Computation of S and ρ.

T	Y	f	$d = Y - T$	d^2
5.94	4	1	−1.94	3.7636
	6	1	0.06	0.0036
15.44	15	1	−0.44	0.1936
	16	2	0.56	0.3136
	17	2	1.56	2.4336
	18	2	2.56	6.5536
	19	1	3.56	12.6736
20.19	15	1	−5.19	26.9361
	16	1	−4.19	17.5561
	17	2	−3.19	10.1761
	18	4	−2.19	4.7961
	19	2	−1.19	1.4161
	20	1	−0.19	0.0361
	21	1	0.81	0.6561
2019	21	1	0.81	0.6561
	22	2	1.81	3.2761
	23	2	2.81	7.8961
	24	2	3.81	14.5161
	25	1	4.81	23.1361
15.44	10	1	−5 44	29 5936
	14	1	−1.44	2.0736
		$n = 32$	$\Sigma fd = -0\ 08$	$\Sigma fd^2 = 229.6252$
			$c = -0.0025$	$S^2 = 7.1758$

$\rho^2 = 1 - S^2/\sigma_y^2 = 1 - 7.1758/21.2773 = 1 - 0.337 = 0.663$

$\rho = 0.814.$

An alternative approximation, $\rho = \sigma_t/\sigma_y$ also gives 0.814.

For purposes of plotting it is best to locate the mode (Mo) and the height at the mode (Y_{mo}) of the parabola, as follows:

$$Mo = - b/2c = - 2.375/-4.75 = 0.5$$

$$Y_{mo} = a - b^2/4c = 20.19 - 5.6406/(-9.5) = 20.784$$

The computation consists of finding, first, the standard deviation of the Y series. This is done by the usual short-cut method of standard deviation in which an assumed average ($M_a = 18$) is taken as the basis of the deviations; the resulting correction ($c = -0.1875$), however, is unimportant. The next step consists of finding the standard error of estimate, that is, the standard deviation of the residuals of the data from the trend. In finding the standard error of estimate each Y is diminished by the T representing the parabolic regression point for that particular column, to obtain the required deviations (d). These deviations, weighted with the frequencies, should total approximately

zero. The deviations are then squared, multiplied by f, and totaled to obtain Σd^2. This total divided by 32 is the square of the standard error of estimate. The value of ρ may next be found directly from the general formula of correlation, as indicated. The result is $\rho = 0.81$, which is somewhat larger than the r obtained indirectly by means of the coefficient of similarity. The difference between the two results is due to irregularities in the data, and probably the smaller of the two results is to be preferred.

Other methods of curvilinear correlation.—There are many complex cases of curvilinear correlation which might be considered, but the complexity inheres in the determination of the regression curve rather than in the determination of the coefficient. Sometimes it is possible to reduce a curvilinear correlation to linear correlation by a change in one or both of the scales, as from an arithmetic to a logarithmic scale. In other cases it may be necessary to fit somewhat complicated regression curves, such as hyperbolas or even growth curves. On account of the fact that it is very difficult to determine the appropriate form of the regression curve, coefficients of curvilinear correlation are really not very dependable, since by chance an inappropriate regression curve may sometimes give a fairly close fit. By an appropriate regression curve is meant one which expresses the general law likely to be effective in the field of which the given case is a sample. This cannot be determined precisely by mathematical methods, but only by a broad study of the nature of the problem itself. Hence if in any given problem a regression line is used which seems to give the closest fit, the degree of correlation may be exaggerated. In any case it is well to plot the regression as a picture of the correlation, and to take the resulting coefficient merely as tentative until enough comparable studies have been made to form a basis for a generalization regarding the nature of the regression. For most purposes, linear, parabolic, and logarithmic regressions will be found sufficient. The use of the means of the columns as a regression curve, as illustrated in the correlation ratio, is not advisable in general.

SUPPLEMENTARY METHODS

Partial correlation.—Two series may correlate, or fail to correlate, largely because of the influence of a third factor upon them, rather than because of their inherent relationship.* A simple case of this sort

* In partial and multiple correlation the dependent series are assumed to register either directly or indirectly the consequences of causes expressed by the independent series. There may, of course, be some cases where mere interrelations may lend themselves plausibly to such correlations, but as a rule, the assumption is as just stated.

arises when two sets of data exhibiting both cyclic and trend variability are correlated without first eliminating the trends. The resulting coefficient may be due partially or wholly to the trend influence, rather than to the cycle. The correlation of the cycles independent of trend influence may of course be found as previously explained by eliminating the trends and correlating the cycles. This process may, however, be combined into an operation called partial correlation, in which trend influence is eliminated by introducing time as a third correlative, and eliminating its influence. The formula is as follows (the symbol $r_{12 \cdot 3}$ means series 1 and 2 correlated after eliminating from each the influence of series 3; r_{12} or r_{21} means the simple correlation of series 1 and 2; etc., all correlations being assumed linear):

$$r_{12 \cdot 3} = (r_{12} - r_{13}\, r_{23}) \div [(1 - r_{13}^2)(1 - r_{23}^2)]^{1/2}$$

The process is illustrated by the accompanying example (Example 77).

Example 77.—Partial correlation of two time series expressed as deviations from their average. Time, centered, is also expressed as deviations, and is eliminated by the process of partial correlation ($r_{xy.t}$) by formulas given below. In effect the process correlates the residuals of x and y after eliminating a trend from each.

Deviations			Products		
x	y	t	xy	xt	yt
−4	12	−3	−48	12	− 36
−2	6	−2	−12	4	− 12
0	6	−1	0	0	− 6
3	3	0	9	0	0
−1	− 7	1	7	−1	− 7
−1	− 9	2	9	−2	− 18
5	−11	3	−55	15	− 33
σ=2.828	σ=8.246	σ=2	−90	28	−112
			r=−0.551	r=0.707	r=−0.970

$$r_{xy.t} = (r_{xy} - r_{xt}\, r_{yt}) \div [(1 - r_{xt}^2)\,(1 - r_{yt}^2)]^{1/2}$$

$$= (-0.551 + 0.707 \times 0.970) \div (0.5000 \times 0.0588)^{1/2} = 0.1348/0.1715 = 0.786$$

It will be seen that the direct correlation of the x- and y-deviations is $r = -\ 0.55$, owing to the fact that the trends have unlike slopes; but with straight-line trend influence removed by the partial correlation against time, the result becomes $r = 0.79$. This result is identical with that obtained by eliminating the trends and correlating the " difference " cycles, $Y - T$; and nearly the same as if the percentage cycles $Y/T - 100$ were correlated. The slopes of the x- and y-cycles are

1 and -4, respectively, and the cycles remaining after eliminating these trend slopes are

x......	-1;	0;	1;	3;	-2;	-3;	2
y......	0;	-2;	2;	3;	-3;	-1;	1

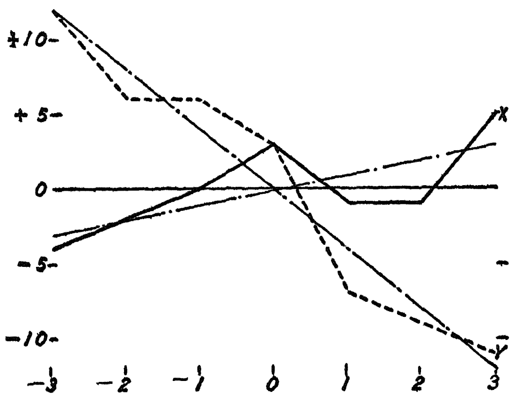


CHART 39

Correlation of two times series, X having an upward trend and Y having a declining trend (cf. Example 77). The direct linear correlation is $r = -0.55$, owing to the opposite trend effects.

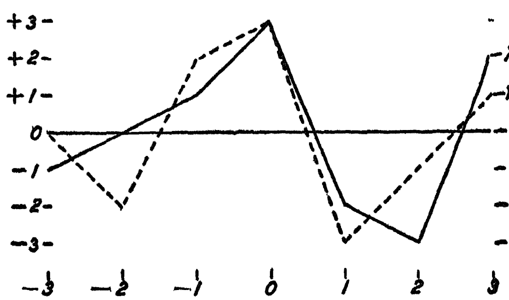


CHART 39a

The same data as in Chart 39 with trends eliminated. The correlation now is $r = 0.79$, which expresses the relationship of the cycles. The elimination of the trends may in effect be accomplished by the use of partial correlation in which t (time) is eliminated.

the correlation of which is

$$r = \Sigma xy/n\sigma_x\sigma_y = 22 \div (7 \times 2 \times 2) = 0.786$$

A partial correlation of two series may be shown to be equivalent to subtracting the third or eliminated series from each of the two

major correlatives, and correlating the residual cycles thus obtained, provided that the data are first expressed in units of the standard deviation from the average, and that in each subtraction the eliminated series is multiplied by the coefficient expressing the correlation between the two series involved in the subtraction. That is, assuming the standard deviation series, x, y, and z; the partial correlation coefficient, $r_{xy.z}$, is the simple correlation of the two series,

$$(x - zr_{xz}) \quad \text{and} \quad (y - zr_{yz})$$

and the partial coefficient may be obtained by the procedure thus indicated. When time is the third series, trend influence is eliminated, since, with σ series, tr_{xt} is the trend of x and tr_{yt} is the trend of y.

Partial correlation may be used in many cases where the correlation between two given series is desired, and where both are materially influenced by a third series, or where the third variate is in effect to be reduced to a constant.

If it is required to find the correlation of two series (1 and 2) after eliminating the common influence of two other series (3 and 4) the formula is

$$r_{12.34} = (r_{12.3} - r_{14.3}r_{24.3}) \div [(1 - r^2_{14.3})(1 - r^2_{24.3})]^{1/2}$$

And if three or more series are to be eliminated the formula is carried out to n terms (n following two or more consecutive integers indicates consecutive integers from the preceding number to and including n, and m similarly to $n - 1$):

$$r_{12.345n} = (r_{12.34m} - r_{1n.34m}r_{2n.34m}) \div [(1 - r^2_{1n.34m})(1 - r^2_{2n.34m})]^{1/2}$$

And in this way formulas of any order may be written. The solutions of the higher order formulas obviously require the prior solution of formulas of a lower order.

Partial correlations may also be obtained by employing multiple correlation, as explained in the following section, using the formula (Rietz, p. 101):

$$1 - r^2_{12.34n} = (1 - r^2_{1.234n}) \div (1 - r^2_{1.34n})$$

Multiple correlation.—A useful application of linear correlation is one which requires the correlation of a given dependent series with several independent series combined. Such a measurement is called a multiple correlation. In effect it requires a grouping of the independent series into a composite series, using such weights as are appropriate to the importance of the several series thus combined. The composite is then correlated with the dependent series. The procedure

is so complex, however, that it may well be expressed by determinants. Let us assume that the correlation of series 1 with series 2, 3, etc., combined (or $r_{1.23n}$) is desired. It is necessary first to find the simple correlations of series 1 and 2 (r_{12} or r_{21}), 1 and 3 (r_{13} or r_{31}), 2 and 3 (r_{23} or r_{32}), etc. Then

$$r^2_{1.234} = 1 - R/R_{11}$$

where

$$R = \begin{vmatrix} 1 & r_{12} & r_{13} & r_{14} \\ r_{21} & 1 & r_{23} & r_{24} \\ r_{31} & r_{32} & 1 & r_{34} \\ r_{41} & r_{42} & r_{43} & 1 \end{vmatrix}$$

and R_{11} is a minor of R formed by dropping column 1 and row 1. The form for more series will be obvious.

The student not familiar with the expansion of determinants may confine his attention to the second method of solution by weights and the Doolittle form given below. However, the solution by determinants may be briefly described as follows:

Designate the first column of the determinant, reading down, as a, b, c, n. And let the minor of a (i.e., R without the column and row in which a appears) be written as A; of b, B; etc. Then,

$$R = aA - bB + cC - \ldots nN \quad \text{(signs alternating)}$$

A determinant is thus reduced to a combination of minors, each of which may be attacked by a repetition of the same process, until the minors take the form

$$R_x = \begin{vmatrix} a & c \\ b & d \end{vmatrix}$$

which is evaluated as $R_x = ad - bc$. The minors may then be successively recombined into majors of higher orders until the original determinant is solved. The process becomes excessively laborious, however, for a very complex multiple correlation. Certain short cuts are available, however, as explained in advanced text-books in algebra or handbooks of engineering.

A second method of computing a multiple correlation will make the nature of the process clearer. If each of the series of deviations to be correlated is reduced to units of its own standard deviation, the independent series may be combined into a single series, which in turn may be correlated with the dependent series by the usual linear method. If appropriate weights can be determined, the independent series may

be combined merely by taking the weighted totals of the correlative items. The weights, in turn, may be found by a method analogous to that employed in finding the parameters of parabola trends. Normal equations are written such that, according to the principle of least squares, the weights required to give the maximum correlation will be determined. It is as if the independent series were to constitute a forecasting index, and the problem was so to combine them that the maximum correlation with the dependent series might be obtained. The normal equations determining the weights have as coefficients of the weights the minor R_{11} previously described. For three series, to be correlated as expressed by the symbol $r_{1.23}$, they are:

$$w_2 + r_{23}w_3 = r_{12}$$

$$r_{32}w_2 + \quad w_3 = r_{13}$$

where w_2 and w_3 are the required weights, to be applied to series 2 and 3, respectively. If series 2 and 3, each expressed in units of its own standard deviation, are each multiplied by its own weight and added, the composite series thus found may be correlated with series 1 to obtain the required coefficient of multiple correlation. In practice, however, this procedure may be very much abbreviated by the use of the formula

$$r^2_{1.23} = r_{12}w_2 + r_{13}w_3$$

The process is illustrated in the accompanying example, which shows both the application of the weights and the abbreviation by use of the above formula.

Example 78.—Multiple correlation, $r_{1.23}$; series 1 with series 2 and 3 combined. The weights, w_2 and w_3, are determined by the normal equations, and are used in combining series 2 and 3, which are then correlated with series 1. The correlation is also measured more directly by the formula as stated. Each series is expressed in units of its own standard deviation. The linear correlations, $r_{12} = 0.74$; $r_{13} = 0.72$; r_{23} (or r_{32}) $= 0.88$ are assumed to have been previously calculated.

I. Calculation of weights, w_2 and w_3. Normal equations:

$$w_2 + r_{23}w_3 = r_{12}; \qquad w_2 + 0.88w_3 = 0.74$$

$$r_{23}w_2 + \quad w_3 = r_{13}; \quad 0.88w_2 + \quad w_3 = 0.72$$

Expressed as determinants:

$$w_2 = \begin{vmatrix} r_{12} & r_{23} \\ r_{13} & 1 \end{vmatrix} \div \begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}; \quad w_3 = \begin{vmatrix} 1 & r_{12} \\ r_{23} & r_{13} \end{vmatrix} \div \begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}$$

Solving:

$$w_2 = (r_{12} - r_{13}\,r_{23}) \div (1 - r^2_{23}) = (0.74 - 0.6336) \div 0.2256 = 0.4716$$

$$w_3 = (r_{13} - r_{12}\,r_{23}) \div (1 - r^2_{23}) = (0\ 72 - 0.6512) \div 0.2256 = 0.3050$$

The solution may be by elimination, or by determinants; in the latter case, w_2 equals a fraction having as its denominator a determinant formed by the coefficients of the first members of the normal equations, and as its numerator the same determinants with the second members of the equations substituted for the coefficients of w_2. The other weight is similarly expressed. This is the method indicated. The resulting equations may be used with other problems of the order $r_{1.23}$.

II. Correlation of series 1 with combined series 2 and 3.

Series (1)	(2)	(3)	(4) (2) × 0.4716	(5) (3) × 0.3050	(6) (4) + (5)	(7) (1) × (6)
1.2	0.0	0.0	0.0	0 0	0.0	0.0
−1.0	−2 0	−1.4	−0 9432	−0.4270	−1.3702	1.3702
−0.8	−0.5	−0.9	−0.2358	−0 2745	−0.5103	0.4082
−0.4	0.0	0.8	0.0	0 2440	0.2440	−0.0976
1.2	1.0	1.1	0.4716	0 3355	0.8071	0.9685
1.4	1.5	1.2	0.7074	0.3660	1.0734	1.5028
−0.6	0.5	0.5	0.2358	0.1525	0.3883	−0.2330
−1 0	−0.5	−1.3	−0.2358	−0.3965	−0.6323	0.6323
σ=1.0					σ=0.7543	4.5514

Series 6 is weighted total of series 2 and 3

$$r_{1.23} = \Sigma(1)\,(6)/n\sigma_1\,\sigma_6 = 4.5514/(8 \times 1 \times 0.7543) = 0.754$$

Abbreviated method, using formula:

$$r^2_{1.23} = r_{12}\,w_2 + r_{13}\,w_3 = 0.74 \times 0.4716 + 0.72 \times 0.3050 = 0.5686$$

$$r_{1.23} = 0.754$$

Multiple correlation, several series.—Weights for combining d/σ independent series ($x_2x_3x_4$) to obtain the maximum correlation with a dependent series (x_1) are obtained from the equations, which may be expanded to n equations (r_{22}, etc., $=1$):

$$r_{22}w_2 + r_{23}w_3 + r_{24}w_4 = r_{12}$$

$$r_{23}w_2 + r_{33}w_3 + r_{34}w_4 = r_{13}$$

$$r_{24}w_2 + r_{34}w_3 + r_{44}w_4 = r_{14}$$

It is not necessary, however, to combine the independent series by

these weights in order to get their combined correlation with the dependent series. The required $r_{1.234}$ is given as

$$r^2_{1.234} = w_2 r_{12} + w_3 r_{13} + w_4 r_{14}$$

which may be expanded for n series.

In extended problems the solution of the weights equations may be facilitated by the use of the Doolittle method, which reduces the amount of calculation and provides a running check (see Example 79). The data lines I, II, III are set down as in the above equations, omitting weights. Under "Operations," lines and factors are indicated in algebraic form. Thus (1)(b, 2) means each item in line 1 times -0.88, located at column b line 2 (begin in column b). The abbreviation "Neg." indicates a change of algebraic sign. The procedure with four independents is suggested at the foot of the table, and the extension with five or more independents will be obvious.

Example 79.—Multiple correlation—Doolittle method. Coefficients of linear correlation, as required in the formulas for multiple correlation (page 270) are assumed as data (cf. Example 49a, p. 166).

Operations (line and column)		a	b	c	y	Check (Σ)	
	I	1.00	0.88	−0.90	0 74	1.72	
	II	(0.88)	1.00	−0.95	0.72	1.65	
	III	(−0.90)	(−0.95)	1.00	−0.88	−1.73	
I	(1)	1.00	0.88	−0.90	0.74	1.72	
Neg. (1)/(a, 1)	(2)	−1.00	−0.88	0.90	−0.74	−1.72	
II	(3)	....	1.00	−0.95	0.72	1.65	
(1) (b, 2)	(4)	....	−0.7744	0.792	−0 6512	−1.5136	
(3)+(4)	(5)	....	0.2256	−0.158	0.0688	0.1364	
Neg. (5)/(b, 5)	(6)	....	−1.00	0.7004	−0.3050	−0.6046	check
III	(7)	....	....	1.00	−0.88	−1.73	
(1) (c, 2)	(8)	....	....	−0.81	0.666	1.548	
(5) (c, 6)	(9)	....	....	−0.1107	0.0482	0.0955	
(7)+(8)+(9)	(10)	....	....	0.0793	−0.1658	0.0865	
Neg. (10)/(c, 10)	(11)	....	....	−1.00	2.0908	1.0908	check
Weights	(12)	−0.1214	−1.1594	−2.0908	(−1.00)		

(y, 12) = −1 (insert)

(c, 12) = (y, 12) (y, 11) = − 2.0908

(b, 12) = (c, 12) (c, 6) + (y, 12) (y, 6) = (−2.0908) (0.7004) +(−1) (−0.3050) = − 1.1594

(a, 12) = (b, 12) (b, 2) + (c, 12) (c, 2) + (y, 12) (y, 2) = (−1.1594) (−0.88) +(−2.0908) (0.90) + (−1) (−0.74) = − 0.1214

$$r^2_{y.abc} = r_{ay} w_a + r_{by} w_b - r_{cy} w_c$$
$$= -0.74 \times 0.1214 - 0.72 \times 1.1594 + 0.88 \times 2.0908 = 0.9153$$

$$r_{y.abc} = 0.957$$

Note: Another independent series would add line IV and column d, and a new section after line 11, with the designations:

IV	(12)
(1) $(d, 2)$	(13)
(5) $(d, 6)$	(14)
(10) $(d, 11)$	(15)
(12)+(13)+(14)+(15)	(16)
Neg. (16)/$(d, 16)$	(17)
Weights	(18)

The weights would then be:

$(y, 18) = -1$ (insert)
$(d, 18) = (y, 18)\ (y, 17)$
$(c, 18) = (d, 18)\ (d, 11) + (y, 18)\ (y, 11)$
$(b, 18) = (c, 18)\ (c, 6) + (d, 18)\ (d, 6) + (y, 18)\ (y, 6)$
$(a, 18) = (b, 18)\ (b, 2) + (c, 18)\ (c, 2) + (d, 18)\ (d, 2) + (y, 18)\ (y, 2)$

Estimation by multiple correlation.—If $r_{1.23n}$ has been found for deviation series $d_1, d_2, d_3 \ldots d_n$, the most probable magnitude of d_1 may be estimated from known magnitudes of d_2, d_3, etc., assuming data for which the computed correlation is appropriate. The estimation equation is

$$d_1/\sigma_1 = (d_2/\sigma_2)w_2 + (d_3/\sigma_3)w_3 + \ldots (d_n/\sigma_n)w_n$$

or

$$d_1 = d_2(\sigma_1/\sigma_2)w_2 + d_3(\sigma_1/\sigma_3)w_3 + \ldots d_n(\sigma_1/\sigma_n)w_n$$

If the deviations are from means, this equation may be written by substituting $d_1 = Y_1 - M_1$ etc., where Y_1 refers to the original series from which the deviations were computed, and M_1 is the average. Any other series may similarly be estimated by regarding it as dependent and recomputing. The process is applied to barometric series as independent, or to any set of interrelated variables where any one may provisionally be taken as dependent. With only two variables, obviously, $w_2 = r_{12}$.

Example 80.—Estimates by means of multiple correlation. The most probable (trend) magnitude of Z, taken as dependent, is estimated from correlative values of X and Y, taken as independent, the following calculations being assumed as preliminary (M = mean; w = weights, as above):

$$M_z = 20;\ M_x = 10;\ M_y = 25$$

$$\sigma_z = 5;\ \sigma_x = 2;\ \sigma_y = 10$$

$$r_{zx} = 0.738;\ r_{zy} = 0.725;\ r_{xy} = 0.881$$

$$w_x = 0.4456;\ w_y = 0.3342;\ r_{z.xy} = 0.756$$

Z	z	X	x	Y	y
26	6	10	0	25	0
15	−5	6	−4	11	−14
16	−4	9	−1	16	− 9
18	−2	10	0	33	8
26	6	12	2	36	11
27	7	13	3	37	12
17	−3	11	1	30	5
15	−5	9	−1	12	−13

Required: the most probable magnitude of Z; given $X = 12$; $Y = 35$.

$$Z = w_x(\sigma_z/\sigma_x)\,(X - M_x) + w_y(\sigma_z/\sigma_y)\,(Y - M_y) + M_z$$

$$= 0.4456\,(5/2)\,(X - 10) + 0.3342\,(5/10)\,(Y - 25) + 20.$$

$$Z = 1.1140X + 0.1671Y + 4.6825$$

$$= 13.3680 + 5.8485 + 4.6825 = 23.8990$$

Curvilinear multiple correlation, curvilinear regressions.—It often happens that the factors involved in multiple correlation are curvilinear rather than linear. For example, if the effects of varying temperatures and rainfall on crop yields are to be determined from relevant and comprehensive data, curvilinear regression curves may be required. That is, low temperatures and excessively high temperatures alike may give low yields, and moderate temperatures may give high yields. The same type of regression may hold for rainfall. At the same time the influence of each independent series on crop yields could not be determined independently on account of a probable relation between the two. In such a case, two methods of solution are available, one strictly mathematical and the other partly mathematical and partly graphic. We may consider the two methods in the order named.

Mathematical solution.—Let us assume that the accompanying partial data (see Example 81) represent two factors (independent series) U and V, influencing a third set of data (dependent series) Y, in much the same way as that described for temperature, rainfall, and crops. The data are reduced to deviation cycles, as indicated by the small letters, u, v, and y, for convenience of calculation and exposition. They will not, however, be reduced to standard or average deviation cycles, though this step might be taken.

Example 81.—Abbreviated data for illustrating method of curvilinear multiple correlation. Independent series U and V; dependent series Y; deviation cycles (from respective means, M), u, v, and y.

Year	U	u	V	v	Y	y
1901	12	2	10	0	28	8
1902	11	1	8	−2	18	− 2
1903	8	−2	12	2	10	−10
1904	10	0	11	1	25	5
1905	9	−1	9	−1	19	− 1
	5)50		5)50		5)100	
	M = 10		M = 10		M = 20	

It is assumed that the regression is parabolic; that is, if the magnitudes of certain constants (a, b, c, d, and e) are known, and if correlation is perfect, then:

$$y = a + bu + cu^2 + dv + ev^2$$

(other independent series might be added, as $fw + gw^2$, etc.), thus expressing the y-deviations as the total effect of the u- and v-deviations acting according to some undetermined parabola. It is possible to avoid the necessity of a separate constant effecting an adjustment for the height of the curve, by centering u^2 and v^2; that is, by expressing them as deviations from their own averages. The centered squares will be distinguished by underscoring. But if the correlation is not perfect, the equation will not give y, but an estimate approximating y, which may be denoted by y'. The equation in its final form, using a with u, etc., therefore reads:

$$au + b\underline{u}^2 + cv + d\underline{v}^2 = y'$$

If it is assumed that y' shall be as close as possible to y, as measured by the least squares or standard deviation criterion ($\Sigma\overline{y - y'}^2$ a minimum) the constants a, b, c, and d may be determined. The normal equations are formed by multiplying the regression equation just given (equated to y and instead of y') by the coefficients of the constants, successively, and summating. The coefficients may be determined from the data, and the solution may be carried out by the Doolittle method, as appears in Example 82.

Example 82.—Determination of parabolic regression curves for the data of Example 81. Equation of curve (y' estimate of y): $au + b\underline{u}^2 + cv + d\underline{v}^2 = y'$, where u^2 and $\underline{v}^2$ are centered ($\underline{u}^2 = u^2 - \Sigma u^2/n$).

Normal equations:

$$a\Sigma u^2 + b\Sigma u\underline{u}^2 + c\Sigma uv + d\Sigma u\underline{v}^2 = \Sigma uy$$

$$a\Sigma u\underline{u}^2 + b\Sigma\underline{u}^2\underline{u}^2 + c\Sigma\underline{u}^2v + d\Sigma\underline{u}^2\underline{v}^2 = \Sigma\underline{u}^2y$$

$$a\Sigma uv + b\Sigma\underline{u}^2v + c\Sigma v^2 + d\Sigma v\underline{v}^2 = \Sigma vy$$

$$a\Sigma u\underline{v}^2 + b\Sigma\underline{u}^2\underline{v}^2 + c\Sigma v\underline{v}^2 + d\Sigma\underline{v}^2\underline{v}^2 = \Sigma\underline{v}^2y$$

Data u, v, and y, centered (cf. Example 81), and computation of summations required in normal equations.

u	v	y	u^2	v^2	$\underline{u}^2$	$\underline{v}^2$	uv	$u\underline{u}^2$	$u\underline{v}^2$	$\underline{u}^2v$	$v\underline{v}^2$	$\underline{u}^2\underline{u}^2$	$\underline{v}^2\underline{v}^2$	$\underline{u}^2\underline{v}^2$	uy	$\underline{u}^2y$	vy	$\underline{v}^2y$
2	0	8	4	0	2	−2	0	4	−4	0	0	4	4	−4	16	16	0	−16
1	−2	−2	1	4	−1	2	−2	−1	2	2	−4	1	4	−2	−2	2	4	−4
−2	2	−10	4	4	2	2	−4	−4	−4	4	4	4	4	4	20	−20	−20	−20
0	1	5	0	1	−2	−1	0	0	0	−2	−1	4	1	2	0	−10	5	−1
−1	−1	−1	1	1	−1	−1	1	1	1	1	1	1	1	1	1	1	1	5
			10	10			−5	0	−5	5	0	14	14	1	35	−11	−10	−44

Normal equations abbreviated by expressing only the computed coefficients and y-products, as indicated by summated cross-multiplication of row and column designations. Solution by the Doolittle method. The abbreviation "Neg." indicates changed algebraic signs.

		$a(u)$	$+ b(\underline{u}^2)$	$+ c(v)$	$+ d(\underline{v}^2)$	$= y$	Check
(u)	(I)	10	0	−5	−5	35	35
$(\underline{u}^2)$	(II)	0	14	5	1	−11	9
(v)	(III)	−5	5	10	0	−10	0
$(\underline{v}^2)$	(IV)	−5	1	0	14	−44	−34
		Solution					
I	(1)	10	0	−5	−5	35	35
Neg. (1)/(a, 1)	(2)	−1	0	0.5	0.5	− 3.5	− 3.5
II	(3)	...	14	5	1	−11	9
(1) (b, 2)	(4)	...	0	0	0	0	0
(3) + (4)	(5)	...	14	5	1	−11	9
Neg. (5)/(b, 5)	(6)	...	−1	−0.3571	−0 0714	0.7857	− 0.6429
III	(7)	...	...	10	0	−10	0
(1) (c, 2)	(8)	...	...	−2.5	−2.5	17.5	17.5
(5) (c, 6)	(9)	...	...	−1.7855	−0.3571	3.9281	− 3.2139
(7) + (8) + (9)	(10)	...	...	5.7145	−2.8571	11.4281	14.2861
Neg. (10)/(c, 10)	(11)	...	...	−1	0.5	− 2	− 2.5
IV	(12)	...	...	....	14	−44	−34
(1) (d, 2)	(13)	...	...	....	−2.5	17.5	17.5
(5) (d, 6)	(14)	...	...	....	−0.0714	0.7857	− 0.6429
(10) (d, 11)	(15)	...	...	....	−1.4286	5.7141	7.1431
(12)+(13)+(14)+(15)	(16)	...	...	....	10	−20	−10
Neg. (16)/(d, 16)	(17)	...	...	....	−1	2	1
Coefficients	(18)	3	−1	1	−2	(− 1)	

The coefficients are calculated as follows:

(y, 18) = − 1 (insert)
(d, 18) = (y, 18) (y, 17)
(c, 18) = (d, 18) (d, 11) + (y, 18) (y, 11)
(b, 18) = (c, 18) (c, 6) + (d, 18) (d, 6) + (y, 18) (y, 6)
(a, 18) = (b, 18) (b, 2) + (c, 18) (c, 2) + (d, 18) (d, 2) + (y, 18) (y, 2)

The normal equations are derived in the usual way. The sum of the deviations squared $\Sigma d^2 = \Sigma(y - y')^2$ is to be a minimum. Substitute for y' its value, $au + b\underline{u}^2 + cv + c\underline{v}^2$, and differentiate with respect to each a, b, c, and d, respectively, equating each derivative to zero. The coefficients of the constants are easily expressed by summated cross multiplications of the equation of the curve as indicated in Example 82.

The regression equation may now be written:

$$3u - u^2 + v - 2\underline{v}^2 = y'$$

which becomes:

$$\begin{array}{lllll}
3u \quad - u^2 + v - 2\underline{v}^2 & = & y' & y & d \\
3(2) \quad - 2 \; + 0 \; -2(-2) & = & 8 & 8 & 0 \\
3(1) \quad + 1 \; - 2 \; -2(2) & = & -\ 2 & -\ 2 & 0 \\
3(-2) - 2 \; + 2 \; -2(2) & = & -10 & -10 & 0 \\
3(0) \quad + 2 \; + 1 \; -2(-1) & = & 5 & 5 & 0 \\
3(-1) + 1 \; - 1 \; -2(-1) & = & -\ 1 & -\ 1 & 0
\end{array}$$

Since $y' = y$, the correlation is obviously $+1.00$. In case d had been significant, however, then the coefficient of correlation might have been obtained as,

$$r^2_{y.uv} = 1 - S_y{}^2/\sigma_y{}^2$$

where S, the standard error of estimate, is the standard deviation of $d = y - y'$, and σ_y is the standard deviation of the dependent series, Y or y.

Estimates by the regression equation.—If the values of u and v are known in a situation like that studied, then an estimate of y may be made by substituting the known values of u and v in the regression equation. For example, if U is rainfall; V, temperature; and Y, crop yields, then in the area studied a rainfall of 10 units, a temperature of 12 units would lead to the following estimate (see Examples 81 and 82):

Regression equation:

$$y' = 3u - \underline{u}^2 + v - 2\underline{v}^2$$

$$u = U - \Sigma U/n = 10 - 10 = 0$$

$$\underline{u}^2 = u^2 - \Sigma u^2/n = 0 - 2 = -2$$

$$v = V - \Sigma V/n = 12 - 10 = 2$$

$$\underline{v}^2 = v^2 - \Sigma v^2/n = 4 - 2 = 2$$

$$y' = 3(0) - (-2) + 2 - 2(2) = 0$$

But y' is an estimate of y, the deviation of crops from their average; that is $y = Y - \Sigma Y/n$, or $0 = Y - 20$, and Y, the actual crop yield as estimated, is 20. The estimate might be made from a chart (see Chart 40).

Needless to say, the regression equation applied to ranges of deviation and geographical areas (or other situations) outside of those actually studied may easily lead to error.

Solution by graphic approximations.—An approximate solution using graphic methods may now be considered. This solution is begun by first assuming linear regression, and afterwards correcting it by successive approximations. The computation of linear regression is illustrated in Example 83. The coefficients a and b are in effect weights for combining the centered series u and v so as to produce a linear series, y', having the least possible deviation from y; that is, $\Sigma(y - y')^2$ is to be a minimum. The solution for a and b is similar to that previously applied to standard deviation series, but is here applied directly to the deviations themselves ($d = u - \Sigma U/n$). The y' series is written as, $y' = au + bv$ (other series might be included, as cw, etc.), and the deviations, $d = y - y'$, are then obtained. For purposes of graphic approximation next to be considered, the series $au + d$ and $bv + d$ are written. The method may be extended to three or more independent series, but becomes very laborious.

Example 83.—Linear regression: two independent series, u and v, and dependent series y, each series centered (see Example 81). Normal equations: $a\Sigma u^2 + b\Sigma uv = \Sigma uy$ and $a\Sigma uv + b\Sigma v^2 = \Sigma vy$, abbreviated for use with the Doolittle method. Computation of y', $d = y - y'$; and the series $au + d$ and $bv + d$ to be used in graphic approximation.

u	v	y	u^2	v^2	uv	uy	vy
2	0	8	4	0	0	16	0
1	−2	− 2	1	4	−2	−2	4
−2	2	−10	4	4	−4	20	−20
0	1	5	0	1	0	0	5
−1	−1	− 1	1	1	1	1	1
			10	10	−5	35	−10

Normal equations abbreviated as summated cross-multiplications of first row and column; the coefficients of a and b, and the y-products being written. Solution by the Doolittle method.

		$a(u)$	$+ b(v)$	$= y$
(u)	I	10	−5	35
(v)	II	−5	10	−10

Solution

I	(1)	10	−5	35	
Neg. (1)/(a, 1)	(2)	−1	0.5	− 3.5	
II	(3)	...	10	−10	
(1) (b, 2)	(4)	...	−2.5	17.5	Coefficients:
3 + 4	(5)	...	7.5	7.5	(y, 7) = − 1 (insert)
Neg. (5)/(b, 5)	(6)	...	−1	− 1	(b, 7) = (y, 7) (y, 6)
Coefficients	(7)	$a = 4$	$b = 1$	(−1)	(a, 7) = (b, 7) (b, 2) + (y, 7) (y, 2)

Computation of y', and series $au + d$ and $bv + d$ for use in graphic approximation

u	v	au	$+bv$	$=$	y'	y	$(y - y')$ d	$au + d$	$bv + d$
2	0	8	0		8	8	0	8	0
1	−2	4	−2		2	− 2	−4	0	−6
−2	2	−8	2		−6	−10	−4	−12	−2
0	1	0	1		1	5	4	4	5
−1	−1	−4	−1		−5	− 1	4	0	3

Graphic approximations.—In Example 83, the values of y were estimated from u and v by the linear regression, $au + bv = y'$; the

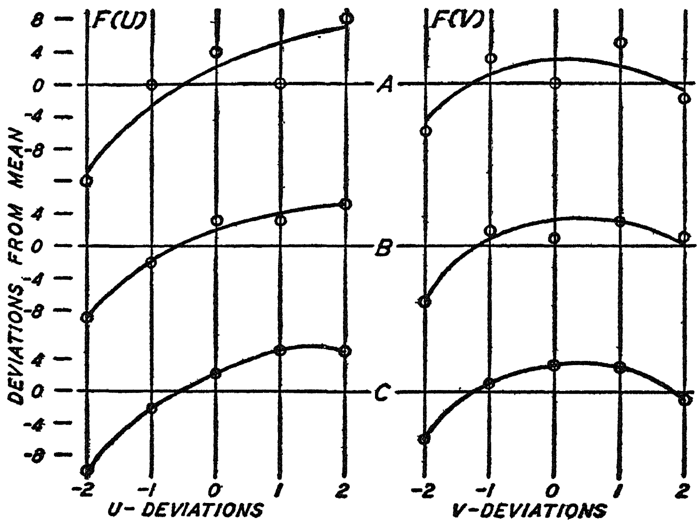


CHART 40

Charts of $(f)u + d$ and $(f)v + d$ where at first (row A) $f(u) = au$ and $f(v) = bv$. The functions are revised by reading from the fitted curves their new values on the u and v ordinates; y' and $d = y - y'$ are recomputed, and $f(u) + d$ is replotted against u, and $f(v) + d$ against v, and new trend lines drawn by inspection (row B). On the basis of the new function curves, $f(u)$ and $f(v)$ are again revised by reading from the chart their new values on the u and v ordinates, and the process of approximating the regression repeated (row C). This regression is taken as final. See Ezekiel, "Methods of Correlation Analysis."

deviations of the actual from the estimated y were written ($d = y - y'$), and were added successively to the au and bv series to obtain $au + d$ and $bv + d$. The purpose of these last series may now be considered. It is evident that $au + bv + d = y$, since $d = y - y' = y - au - bv$. If bv is held constant at its average (zero), then the expression $au + d$ will reflect the influence of u alone; and if plotted against u will give a partial indication of the real regression of y on u. Similarly au may be held constant and $bv + d$ be plotted against v. These two regressions appear in the first row of Chart 40. A smoothed curve (solid line)

is fitted to the plotted points by inspection. If three or more independent series were present, each would be plotted in the same way; as $au + d$ against u; $bv + d$ against v, $cw + d$ against w, etc. With the usual number of items necessary for a valid correlation, usually 25 or more, the series $au + d$, etc., may be grouped at convenient intervals across the charts and averaged, so as to make the probable position of the trend line clearer. This line should center; that is, its sum at the given ordinates should be zero.

The trend or individual regression curves thus determined are read from the chart at the u, etc., ordinates. These figures are the first approximations of the curvilinear regressions. The regression on u is labeled $f(u)$, and on v, $f(v)$, and these new function series replace the au and bv linear functions. It is not necessary, however, to express the new functions by an equation. By adding the new function series, as read from the curves, a revised estimate, y', is obtained: $f(u) + f(v) = y'$.

Example 84.—Graphic approximations to curvilinear regression based on the estimate in Example 83 (see Chart 40).

First approximation, $f(u)$ and $f(v)$, and data for second

u	v	y	$f(u)$	$+ f(v)$	$= y'$	$y - y'$ d	$f(u) + d$	$f(v) + d$
2	0	8	7	3	10	−2	5	1
1	−2	−2	5	−5	0	−2	3	−7
−2	2	−10	−11	−1	−12	2	−9	1
0	1	5	2	2	4	1	3	3
−1	−1	−1	−3	1	−2	1	−2	2

Second approximation, $f(u)$ and $f(v)$, and data for third

u	v	y	$f(u)$	$+ f(v)$	$= y'$	$(y - y')$ d	$f(u) + d$	$f(v) + d$
2	0	8	5	3	8	0	5	3
1	−2	−2	4	−7	−3	1	5	−6
−2	2	−10	−9	0	−9	−1	−10	−1
0	1	5	2	3	5	0	2	3
−1	−1	−1	−2	1	−1	0	−2	1

Third approximation, $f(u)$ and $f(v)$, final regression line, and standard error of estimate (S).

u	v	y	$f(u)$	$+ f(v)$	$= y'$	$(y - y')$ d	d^2	
2	0	8	5	3	8	0	0	$S^2 = \Sigma d^2/n = 2/5$
1	−2	−2	5	−6	−1	−1	1	
−2	2	−10	−10	−1	−11	1	1	$\sigma_y^2 = \Sigma y^2/n = 38.8$
0	1	5	2	3	5	0	0	
−1	−1	−1	−2	1	−1	0	0	

On the basis of the revised function series and y', the process of

approximation is now repeated and second approximations of $f(u)$, $f(v)$, and y' are obtained. This process may be repeated as many times as is necessary to give smoothed regression curves. In the case illustrated (Example 84 and Chart 40), the third approximation is made identical with the plotted points and is taken as the final regression curve.

The charts showing the final approximation to curvilinear regression (row C, Chart 40) present the probable relation of the dependent series to each independent series in turn. Assuming that the data represent crop yields (y) against rainfall (u) and temperature (v), it is seen that in both cases there is a tendency for crops to increase with the given factor, taken alone, up to a certain point, after which an excess produces negative effects. With adequate data the method just illustrated will usually give fairly good approximations, though with rather erratic data it may fail.* It has the advantage, however, that it does not assume regression curves of any particular type, hence is more flexible than the strict mathematical method.

The coefficient of curvilinear multiple correlation, and estimates of y, may be computed in much the same way as has been explained in connection with the mathematical method previously discussed. The coefficient is

$$r_{y.uv} = (1 - S_y^2/\sigma_y^2)^{1/2} = (1 - 0.4/38.8)^{1/2} = 0.99$$

In making estimates of y from given values of u and v, the functions of u and v may be read from the charts of the final regression curves of $f(u)$ and $f(v)$ on u and v, respectively; or they may be read from the table of u and v and their functions, by interpolation, if necessary. Thus, if $U = 10$ and $V = 12$, $u = 0$ and $v = 2$ (see Example 81), by Chart 40, $f(u) = 2$ and $f(v) = -1$. Then,

$$y' = f(u) + f(v) = 2 - 1 = 1, \text{ estimate of } y$$

$$y = Y - \Sigma Y/n;\ 1 = Y - 20;\ Y = 21, \text{ as estimated}$$

* It will be found in some problems that the graphic method, as here described, fails to converge in successive approximations, but instead varies from one extreme to another. This is particularly likely to be the case when only two independent variables are used. In such a case, convergence may be obtained by averaging successive estimates of the functions of the independent variables, using this average as the new estimated function. This may readily be done by writing the series to be plotted against u as $\frac{1}{2}(y - fv + fu)$, and for plotting against v, $\frac{1}{2}(y - fu + fv)$. When these figures are plotted, the regression curves expressing $f(u)$ and $f(v)$, respectively, may be drawn in as a smooth trend line as previously described. The graphic process is by no means exact and may require other adjustments according to circumstances.

EXERCISES

1. Correlate the average deviation cycles, obtained in Problem 8, p. 222, and Problems 9, 10, 11, p. 222, with the following index of business activity in the United States, as derived from the *Annalist* index. Find the coefficient of similarity (Sm), and the coefficient of correlation (r) as derived from it.

Quarter	1923	1924	1925	1926	1927	1928	1929	1930
1	0.62	−0.14	−0.13	0.31	0.83	−0.02	1.53	−0.60
2	1.64	−2.34	−0.38	0.26	0.83	0.23	2.42	−0.97
3	0.55	−2.69	−0.35	0.76	0.32	0.65	2.33	−2.42
4	−0.52	−0.97	0.35	0.95	−0.76	1.11	0.30	−3.72

The foregoing cycle is based upon a trend fitted to the eight years as given, in order to provide a basis for correlation with other cycles similarly derived. The cycle as given in the *Annalist* is as follows:

Quarter	1923	1924	1925	1926	1927	1928	1929	1930
1	1.55	0.60	0.36	0.50	0.72	−0.31	0.85	−1.34
2	2.41	−1.45	0.07	0.40	0.65	−0.14	1.59	−1.74
3	1.36	−1.84	0.03	0.79	0.13	0.18	1.44	−3.11
4	0.33	−0.34	0.61	0.89	−0.91	0.53	−0.46	−4.36

2. Find Pearson's coefficient of correlation for the following untabulated series. Also find the same coefficient by the method of rank differences. Why should the results by these two methods be alike in these problems?

(a) x	y	(b) x	y	(c) x	y	(d) x	y	(e) x	y	(f) x	y	(g) x	y	(h) x	y
1	3	1	1	1	5	1	2	1	4	1	1	1	1	1	3
2	1	2	2	2	4	2	3	2	3	2	3	2	4	2	1
3	2	3	3	3	3	3	4	3	1	3	2	3	5	3	4
4	5	4	4	4	2	4	5	4	2	4	5	4	2	4	2
5	4	5	5	5	1	5	1	5	5	5	4	5	3	5	5

3. Correlate the untabulated series y and x. How can you tell by an inspection of these data what the answer should be?

y	x
600	−1800
−500	1500
−400	1200
−200	600
600	−1800
700	−2100
−300	900
−500	1500

4. Correlate these two series by Pearson's (r) method and by rank differences. Why do the two methods give like results with these data?

(a) X	Y		Supply and Demand (b) bu. X	¢ Y
10	40		400	50
20	60		200	60
30	70		700	20
40	50		100	70
50	10		500	40
60	30		300	30
70	20		600	10

5. The following columns X and Y are assumed to represent measures of related economic conditions in seven given cities. Find the Pearsonian coefficient of correlation (r) and also compute the correlation (r_r) by the ranking method. In this case the result by the two methods should be identical. Why?

Cities	X	Y
A	8	50
B	11	48
C	7	46
D	12	54
E	13	56
F	10	52
G	9	44

6. In these eight cities find the coefficient of correlation between the assumed population and accident rate.

City	Population (thousands)	Accident rate (per million)
A	10	32
B	20	20
C	30	24
D	40	36
E	50	40
F	60	28
G	70	48
H	80	44

7. Find the coefficient of correlation (r) between sanitation and infant mortality in the indexes of these eight cities.

City	Sanitation (X)	Infant mortality (Y)
A	100	98
B	86	108
C	91	104
D	108	98
E	111	94
F	112	90
G	105	100
H	87	108

8. The following figures are the relative international athletic scores in Olympic games 1920 and 1924. Find Pearson's r between these scores and the mean annual temperatures of the country.

	°C.	Score
Belgium	9.1	35
Denmark	7.2	67
Esthonia	4.4	40
Finland	3.1	148
Holland	8.7	26
Norway	3.8	171
Sweden	5.1	123

9. The following figures are the rankings of corn production and price in the years 1903–1913. Find Pearson's r from these rankings, and also apply the method of rank differences.

Year	Production	Price
1903	10	6
1904	7	7
1905	4	8
1906	2	9
1907	6	5
1908	5	1
1909	8	3
1910	3	10
1911	9	4
1912	1	11
1913	11	2

10. From the σ cycles of the production and price of iron for the ten years 1903–1913, find the coefficient of correlation.

Year	Production σ	Price σ
1903	−0.06	+0.12
1904	−0.85	−1.56
1905	+0.79	−0.32
1906	+1.06	+0.16
1907	+0.94	+2.79
1908	−2.40	−0.20
1909	+0.24	+0.12
1910	+0.37	−0.16
1911	−1.03	−0.72
1912	+0.43	−0.16
1913	+0.46	−0.12

11. Find the r of the per capita income (X) and urbanization (Y) of these states by the rank difference method.

	X	Y		X	Y
Maine..................	10	10	Pennsylvania............	8	6
New Hampshire..........	9	7	Delaware................	2	9
Vermont................	11	12	Maryland................	7	8
Massachusetts...........	3	2	Virginia..................	13	13
Rhode Island............	5	1	North Carolina...........	16	15
Connecticut.............	6	5	South Carolina...........	12	16
New York...............	1	3	Georgia..................	15	14
New Jersey..............	4	4	Florida..................	14	11

12. Correlate the relative international athletic scores in Olympic games, 1920 and 1924, with the mean annual temperature of the country represented.

	°C.	Score		°C.	Score
Australia...........	17.2	7.1	Holland............	8.7	26.0
Belgium............	9.1	35.0	Italy...............	15.2	9.6
Czecho-Slovakia....	8.3	2.9	Norway............	3.8	171.0
Denmark...........	7.2	67.0	South Africa........	16.4	6.6
Esthonia...........	4.4	40.0	Spain..............	13 6	1.6
Finland............	3.1	148.0	Sweden............	5 1	123.0
France.............	11.7	16.0	Switzerland.........	8 6	18.0
Great Britain.......	9.6	13.0	United States.......	10.6	11.7

13. The following data are the assessment ratio and value of rural farm properties in Iowa. Calculate the coefficient of similarity and derive from it an estimate of r.

Assessment ratio (%)—assessed value to market value

Value per acre	$15\frac{1}{2}$	$25\frac{1}{2}$	$35\frac{1}{2}$	$45\frac{1}{2}$	$55\frac{1}{2}$	$65\frac{1}{2}$	$75\frac{1}{2}$	$85\frac{1}{2}$	$95\frac{1}{2}$	$105\frac{1}{2}$	$115\frac{1}{2}$	$125\frac{1}{2}$	$135\frac{1}{2}$
300	..	17	4	1	2	1							
260	1	15	56	3	3	0							
220	2	28	251	65	1	0	1						
180	1	7	147	277	34	8	0	1	0	0	0	0	1
140	1	6	65	301	295	71	6	2	0	0	1		
100	..	3	19	53	128	109	55	13	1	0	0		
60	..	1	1	10	14	19	11	8	4	1	1		

14. From the following data find the coefficient of correlation between the merchandise export balance and the gold export balance, and make a graphic comparison with the rest.

Comparison of American foreign trade, gold movements, and price changes

Year	A Cycles in export balance, units of deviation	B Gold export (+) and import (−) balance, millions of dollars	C Ratio of English to American wholesale prices, thousands	D Cycle of American wholesale prices, units of deviation
1880	−37	− 77.1	950	6.08
1881	38	− 97.5	943	5.30
1882	−50	− 1 8	916	9.51
1883	−13	− 6.1	945	5 72
1884	−10	+ 18.3	972	−2.07
1885	55	− 18.2	985	−5.85
1886	−10	+ 22.2	967	−5 64
1887	−12	− 33.2	953	−3.43
1888	−30	− 25.6	948	1.78
1889	−10	+ 49.7	1009	0.99
1890	10	+ 4.3	1046	0 21
1891	−19	+ 68.1	1032	3 42
1892	35	+ 0.5	1052	−0 37
1893	−62	+ 87.5	1038	2.84
1894	71	+ 4.5	1074	−2.95
1895	−47	+ 30.1	1042	0 27
1896	−73	+ 78.9	1088	−0 68
1897	−19	− 44.7	1089	−2 42
1898	85	−105.0	1091	−3 08
1899	8	− 51.4	1081	−0.62
1900	−21	+ 3.7	1102	2.92
1901	32	− 12.9	1042	−0.42
1902	−10	− 3.5	955	3 32
1903	−33	+ 2.1	955	1.18
1904	9	− 17.6	958	0.12
1905	−11	+ 38.9	996	−2.82
1906	3	− 57.6	1030	−1.68
1907	−17	− 63.1	1000	2.58
1908	62	− 75.9	944	−2.08
1909	−16	+ 47.5	898	2.38
1910	−65	+ 75.2	927	2.92
1911	16	− 51.1	992	−2.42
1912	20	+ 8.4	990	2.32
1913	1	+ 8.6	1000	0.18
1914	−75	+ 45.5	1000	−0.88

15. (a) Find the coefficient of similarity (Sm) and the similarity coefficient of curvilinear correlation (η_s), assuming as a regression line the means of the columns. Also find r and η.

(1)

		X-series			
		5	10	15	20
Y-series	14	..	..	1	1
	10	..	1	2	2
	6	1	5	3	
	2	1	3		

(2)

		X-series					
	..	2	4	6	8	10	12
Y-series	40	..	..	1	1	1	
	30	..	1	3	0	3	1
	20	..	0	1	..	1	3
	10	2	1	..	..	..	1

(b) From the following double frequency tables, calculate r by the formula

$$r = (\sigma_x^2 + \sigma_y^2 - \sigma_w^2) \div (2\sigma_x\sigma_y)$$

where σ_w is the quadratic mean of the upper left to lower right slant deviations, taken as of unit class intervals. Find also σ_v^2 (upper right to lower left deviations), and check by the equation

$$\sigma_v^2 + \sigma_w^2 = 2\sigma_x + 2\sigma_y$$

(1)

Y					
3			1	2	1
2		2	4	2	
1	1	2	1		
X	1	2	3	4	5

(2)

Y					
3	1	2	1		
2		2	4	2	
1			1	2	1
X	1	2	3	4	5

(3)

Y					
5	1	1			
4	1	1	2		
3		2	3	1	
2			2	4	
1				1	1
X	1	2	3	4	5

(4)

Y				
4			1	1
3		1	2	2
2	1	5	3	
1	1	3		
X	1	2	3	4

(5)

Y						
4			1	1	1	
3		1	3	0	3	1
2		0	1		1	3
1	2	1				1
X	1	2	3	4	5	6

16. The following figures are measures applied to certain states. Find the multiple correlation of density, schools, and capital combined as compared with notables. For explanation of terms, see Reinhardt and Davies, "Principles and Methods of Sociology," Appendix.

	x	a	b	c	
x	1.00	0.73	0.87	0.89	x notables
a	....	1.00	0.51	0.93	a density
b	....	....	1.00	0.67	b schools
c	....	....	....	1.00	c capital

17. The correlation between the foreign born (b) and urbanization (c) and the native birth rate (a) are given below. Find $r_{1\cdot 23}$. Subscripts 1, 2, and 3 refer to a, b, and c, respectively.

	a	b	c
a	1.000	−0 800	−0.819
b	−0.800	1.000	0.694
c	−0.819	0 694	1 000

18. The following figures are interest rates and wholesale prices (1909–1913 quarterly). Find r_{12}, r_{13}, and r_{23} by the formula $r = Sm(2 - \Sigma's'/n)$ and the partial correlation $r_{12.3}$.

Year	(1) Interest	(2) Prices	(3) Time	Year	(1) Interest	(2) Prices	(3) Time
1909	−0 29	−2.11	−1.90	1911	−1.47	−0.88	0 10
	−0 20	−1.75	−1.70		−2.06	−0.35	0.30
	−0.59	−0.70	−1.50	1912	−1.47	−0.18	0.50
	1.18	1.40	−1 30		−0.98	1.05	0.70
1910	1.27	1.93	−1.10		−0.20	1.23	0.90
	1.37	0.88	−0.90		0.49	1.93	1 10
	1.76	0.35	−0.70	1913	0.98	1.23	1 30
	0 29	−0.18	−0.50		1.67	−0.18	1.50
1911	−0.59	−0.88	−0 30		0.98	−0.53	1.70
	−1.86	−1.93	−0.10		−0.29	−0.35	1.90

19. From the following ranking of specified states in 1860, find the partial correlation $r_{12.34}$, $r_{13.24}$, $r_{14.23}$, and the multiple correlations $r_{1.234}$, $r_{1.2345}$. (cf. Exercise 16).

	Notables (1)	Density (2)	Education (3)	Capital (4)	Coolness (5)
Alabama..........	24	20	23	24	25
Arkansas..........	29	28	27	29	24
Connecticut.......	2	3	2	3	9½
Delaware..........	8	9	19	8	16
Florida...........	27	29	29	28	29
Georgia...........	25	21	25	23	23
Illinois............	14	13	14	16	11
Indiana...........	17	10	16	14	14
Iowa.............	16	27	12	26	6
Kentucky.........	20	14	22	15	20
Louisiana.........	26	24	20	25	28
Maine............	6	18	4	12	4
Maryland.........	13	6	15	9	18
Massachusetts.....	1	2	1	2	7
Michigan..........	11	26	9	17	5
Mississippi........	28	23	17	27	26
Missouri..........	19	22	18	19	19
New Hampshire....	5	11	5	7	3
New Jersey........	9	4	13	4	13
New York.........	7	5	7	6	8
North Carolina....	22	19	28	22	17
Ohio.............	10	8	11	10	12
Pennsylvania......	12	7	10	5	15
Rhode Island......	4	1	8	1	9½
South Carolina.....	21	17	21	21	27
Tennessee.........	23	15	24	18	22
Vermont..........	3	12	3	11	1
Virginia...........	18	16	26	13	21
Wisconsin.........	15	25	6	20	2

20. From the following deviations, find the correlations r_{12}, r_{13}, r_{14}, r_{23}, r_{24}, r_{34}, $r_{12.3}$, $r_{13.2}$, $r_{14.3}$, $r_{42.3}$, $r_{14.23}$.

Cities	(1)	(2)	(3)	(4)	(5)	(6)
A	3	2	0	6	0	1
B	−4	−6	−14	−5	−4	−4
C	−2	−2	− 9	−4	−1	−2
D	1	0	8	−2	0	1
E	3	2	11	6	2	3
F	5	8	12	7	3	5
G	3	0	5	−3	1	0
H	−4	−4	−13	−5	−1	−4

21. From the following figures, compute the partial correlations $r_{12.3}$, $r_{13.2}$, $r_{14.2}$, $r_{34.2}$, $r_{14.23}$, and the multiple correlations $r_{1.23}$, $r_{1.234}$.

Dependent (effect)		Independent (causes)	
(1)	(2)	(3)	(4)
26	10	25	11
15	6	11	16
16	9	16	14
18	10	33	11
26	12	36	9
27	13	37	7
17	11	30	12
15	9	12	16

22. The following figures are stocks, business activity, and interest rates for eight years by 6-month periods. Compute the partial correlations $r_{12.t}$, $r_{13.t}$, $r_{23.t}$, r_{12}, where t = time. Find $r_{3.12}$ by the Doolittle method.

Allow for a lag of six months in (2) and one year in (3) as indicated below.

Year (t)	Stocks (1)	Business Activity (2)	Interest (3)
			2
		6	7
1903	2	2	6
	− 8	− 8	−10
1904	−14	−13	−13
	− 9	− 3	− 8
1905	5	− 2	3
	9	6	7
1906	16	8	11
	12	15	12
1907	10	14	11
	− 5	− 8	− 9
1908	−20	−17	−18
	− 6	− 9	−13
1909	4	− 2	3
	11	11	10
1910	10	4	
	− 2		

23. By the quadratic parabola curvilinear method and also by the method of approximations based on linear correlation, find a coefficient of correlation.

Years	U	V	Y
1910	24	20	56
1911	22	16	36
1912	16	24	20
1913	20	22	50
1914	18	18	38

24. Assuming that the following functions of the independent variables u and v have been estimated by the graphic method of curvilinear multiple correlation, and that the dependent variable (y) is as given, obtain by further graphic procedure a more accurate measure of the same functions, and compute $r_{y.uv}$.

u	$f(u)$	v	$f(v)$	y
−2	7	4	0	5
1	− 1	−6	−9	−10
3	−11	0	6	− 1
−3	1	−4	2	1
2	− 3	2	5	2
−1	6	−2	5	7
0	1	6	−9	− 4

25. Given the following independent variables, u and v, and the dependent variable y, calculate by parabolic functions the correlation of y on u and v. Also estimate the same correlation by the graphic method. See footnote, p. 280.

Year	u	v	y
1911	−2	1	5
1912	1	−2	− 4
1913	2	3	4
1914	−3	−1	− 2
1915	0	0	4
1916	3	−3	−14
1917	−1	2	7

26. From the following rankings for various phenomena for the Atlantic Coast states, compute correlations by the ranking method.

State	A	B	C	D	E
Maine................	11	11	6	9	6
New Hampshire........	13	13	10	14	8
Vermont..............	14	14	11	2	5
Massachusetts.........	3	4	13	$11\frac{1}{2}$	14
Rhode Island..........	12	9	8	6	15
Connecticut...........	10	5	12	$7\frac{1}{2}$	11
New York.............	1	1	14	10	13
New Jersey............	4	3	15	5	12
Pennsylvania..........	2	2	9	$11\frac{1}{2}$	10
Delaware..............	15	15	2	*	7
Maryland..............	9	7	3	1	9
Virginia...............	7	8	$4\frac{1}{2}$	3	4
North Carolina........	5	6	$4\frac{1}{2}$	$7\frac{1}{2}$	2
South Carolina........	8	12	1	13	1
Georgia...............	6	10	7	4	3

* No state prison or reformatory in Delaware.

A. Population in 1930.

B. Value added by manufacture in 1929.

C. Death rate per 1000 infants under one year of age, 1929.

D. Prisoners received from courts per 100,000 population in state and federal prisons, 1928.

E. Percentage of population rural, 1930.

Care should be exercised in attempting to attach significance to the correlations which are worked out. The intent is only that of providing simple problems to promote familiarity with the method.

27. Correlate the quarterly data for total foreign capital issues and total exports of the United States for the years 1923–1931. The data can be found on pp. 92 and 106 of the *Survey of Current Business*, Annual Supplement, 1932. The data are given by months, but to simplify the computations the monthly figures may be combined to form quarterly data.

ANSWERS

1. $Sm = 0.491$; $r = 0.651$ (from Sm)

2. (a) $r = 0.60$ (c) $r = -1.00$ (e) $r = 0.10$ (g) $r = 0.20$
(b) $r = 1.00$ (d) $r = 0.00$ (f) $r = 0.80$ (h) $r = 0.50$

3. $r = -1$

4. (a) $r = -0.64$ (b) $r = -0.86$

5. $r = r_r = 0.79$ **7.** $r = -0.94$ **9.** $r_r = -0.72$ **11.** $r_r = 0.83$

6. $r = 0.71$ **8.** $r = -0.78$ **10.** $r = 0.45$ **12.** $r = -0.75$

13. $Sm = 0.529$; $Smr = 0.694$ **14.** $r = 0.57$

15. (a)

	(1)	(2)
Sm	0.51	0.09
η_s	0.51	0.57
r	0.683	0.184
η	0.699	0.761

(b)	σ_v^2	σ_w^2	σ_x^2	σ_y^2	r
(1)	2.5	0.5	1.0	0.5	0.707
(2)	0.5	2.5	1.0	0.5	−0.707
(3)	0.49	4.29	1.1	1.29	0.798
(4)	2.5875	0.4875	0.75	0.7875	0.683
(5)	4.35	3.15	2.8	0.95	0.184

16. $r_{x \cdot abc} = 0.9713$

17. $r_{1.23} = 0.880$

18. $r_{12} = 0.5724$; $r_{13} = -0.0587$; $r_{23} = 0.2294$; $r_{12.3} = 0.6030$

19. $r_{12.34} = -0.322$; $r_{13.24} = 0.748$; $r_{14.23} = 0.680$; $r_{1.234} = 0.964$

20. $r_{12} = -0.48$; $r_{13} = -0.70$; $r_{14} = -0.49$; $r_{23} = 0.37$; $r_{24} = 0.36$; $r_{34} = 0.29$; $r_{12 \cdot 3} = -0.33$; $r_{13.2} = -0.636$; $r_{14.3} = -0.43$; $r_{42.3} = 0.28$; $r_{14.23} = -0.38$

21. $r_{12} = 0.7375$; $r_{13} = 0.7250$; $r_{14} = -0.8750$; $r_{23} = 0.88125$; $r_{24} = -0.8958$; $r_{34} = -0.9458$; $r_{12.3} = 0.304$; $r_{13.2} = 0.235$; $r_{14.2} = -0.713$; $r_{34.2} = -0.745$; $r_{14.23} = -0.830$; $r_{1.23} = \pm 0.755$; $r_{1.234} = \pm 0.930$.

22. $r_{12.t} = 0.911+$ (trend removed)
$r_{13.t} = 0.959+$ (trend removed)
$r_{21.t} = 0.956+$ (trend removed)
$r_{12} = 0.92-$
$r_{3.12} = 0.979$

23. $\rho_{y \cdot uv} = 1$

24. $\Sigma' d' =$ nearly 0; $R_{y \cdot uv} =$ nearly 100

25. $\Sigma' d' =$ nearly 0; $R_{y \cdot uv} =$ nearly 100

CHAPTER IX

PROBABILITY AND FREQUENCY CURVES

THROUGHOUT earlier chapters it has been intimated that the statistician works, as a rule, with samples rather than with complete data. Hence questions regarding the probable validity of a sample, and the type of distribution which it follows, continually arise. Theories regarding the probable validity of samples under laboratory conditions have been developed, as, for example, in the flipping of coins, the throwing of dice, or the drawings of diverse units from urns. It was noted earlier, however, that these laboratory studies of probability may not always be applicable to the actual data of the social sciences. One cause of their inapplicability is that in laboratory studies the "observations" are usually independent of each other, whereas in actual social life they are commonly related. For example, when several coins are flipped, each coin falls heads or tails, independently of every other coin. But when several states are studied as economic or social units there is an overlapping of causal factors from one to another which renders boundaries quite artificial. And in the analysis of time series the interdependence of successive months or years is even more obvious. Further, in obtaining samples from data it is easier to control laboratory conditions so that the sample is without bias, but in actual field work it is very difficult to eliminate bias. For, as was pointed out in an earlier chapter, straw votes gathered by publications may be biased because readers of the publication belong, more or less, to certain geographical areas or social classes. Or, again, if random samples are taken from a city telephone directory, the very poor are not adequately represented in the sample. Many other forms of bias in the gathering of samples of social data may be distinguished. Nevertheless, in spite of all the differences between probabilities in the laboratory and in field work, the former studies are of material value as a theoretical background for applied statistics. They belong, however, in their more complex form to advanced statistics rather than to an introductory course; hence in this chapter only an elementary summary is attempted.

The binomial description of probability.—It has already been noted that the binomial expression $(a + b)^s$ will give frequencies which

express the relative probabilities involved in the flipping of a set (s) of coins, or in similar operations of chance. The point binomial thus described may be used in a variety of ways to express probability. To begin with, let us assume a case in which the probability of an event occurring is a given fraction p, then the probability of its not occurring will obviously be $(1 - p)$, which may be denoted as q; or p and q as thus used may be interchanged. Hence, in flipping a coin, where the probability of throwing heads is $\frac{1}{2}$, the probability of not throwing heads—that is, of throwing tails—is also $\frac{1}{2}$. If then a set of 4 coins are thrown simultaneously, the probabilities of throwing 0, 1, 2, 3, or 4 heads may be expressed by the terms of the binomial:

$$(1/2 + 1/2)^4 = (1 + 4 + 6 + 4 + 1) \div 16$$

in which the successive frequencies of the expanded binomial give respectively the relative probabilities of throwing 0, 1, 2, 3, or 4 heads (or tails), respectively. If 16 successive throws are made, the theoretical distribution of the number of heads is given by the following tabulation expressed in class marks (m) and frequencies (f):

$$m = 0, 1, 2, 3, 4 \text{ (heads)}$$

$$f = 1, 4, 6, 4, 1 \text{ (throws)}$$

If, however, dice are thrown instead of coins and the probabilities of throwing aces are tabulated, the odds become 1 in 6 for each die. The probability of not throwing an ace is obviously $\frac{5}{6}$; hence if q represents the probability of throwing an ace $(p + q) = (5/6 + 1/6)$. Supposing now that a set of two dice are thrown simultaneously several times, and the number of aces counted, the probabilities 0, 1, or 2 are

$$(5/6 + 1/6)^2 = (25 + 10 + 1) \div 36$$

that is, there are 25 chances out of 36 of throwing no aces; 10 chances out of 36 of throwing one ace only; and 1 chance out of 36 of throwing two aces. Similarly, the probabilities of throwing aces in larger sets of dice are expressed in general as $(5/6 + 1/6)^s$.

Probability formulas express only what is most likely to occur, not what must occur; and there are different degrees of likelihood. It is easily seen that the probability of approximating the theoretical distribution increases with the number of throws, to an extent that will be suggested later. That is, if sets of coins or dice are thrown only a few times, it is not very likely that the theoretical probability distribution will be realized. But if many throws (n) are made and the results tabulated—that is, if n is increased—the likelihood that the actual

frequency tabulation will approach the theoretical distribution is gradually increased. This tendency toward normality is an illustration of the statistical stability usually inherent in large numbers—a principle which is generally recognized as the basis of actuarial science. It should also be observed that when s as well as n is increased, the expression $(1/2 + 1/2)^s$ approaches the normal probability curve and the characteristics of the distribution may be assumed to be described by tables of that curve (ordinates and integral). In general, the binomial $n\,(p + q)^s$, where both s and n are large, may be assumed to express the logarithmic normal distribution, provided that the scale of class marks (m = 0, 1, 2, etc.) is taken as a logarithm on any convenient base, as e^m, which gives successive values of m (or x) as 1, 2.718, 8.389, 20.086, etc. ($e^{-x^2/2}$ will make the logarithmic standard deviation unity). An examination of this type of distribution shows that if plotted on double-log paper, or on ordinary ratio paper at equal x-intervals, it approximates a parabola.

For purposes of analyzing binomial distributions of the $(p + q)^s$ type, where the class marks, or x-scale, are expressed by the series 0, 1, 2, 3, etc., the following mathematical characteristics may be briefly noted. The average (AM) and standard deviation (σ) of such a distribution are expressed by the formulas:

$$AM = sq; \qquad \sigma = \sqrt{spq}$$

also the number of classes is $(s + 1)$, and the sum of the frequencies equals their common denominator, which is the common denominator of the sth powers of p and q. The mode (Mo), if the frequencies are assumed to be spread over a unit class interval, may be estimated as

$$Mo = [(s + 1) \div (r + 1)] - 0.5$$

where $r = p/q$; or, more exactly, as the positive integer lying between $sq - p$ and $sq + p$.

From these relations it follows that if there is given a total group of items which may be assumed to follow a symmetrical binomial distribution, or a logarithmic distribution that can be reduced to an approximation of the normal type, it is possible to estimate various measures of this distribution. For example, if n is given, s is derived as $n = 2^s$; $s \log 2 = \log n$; and $s = \log n/\log 2$; from which the arithmetic mean and standard deviation may be determined. The position of any given item in the distribution may then be readily fixed. For example, the item (U) at the upper end of the magnitude scale marking the extreme upper range may be located as $[(s/2) \div (\sqrt{s}/2)]$ standard deviation units from the center, or $\sqrt{s}$. The skewed binomial

distribution may be determined in much the same way if an appropriate measure of the skewness is given.

Permutations and combinations.—Binomial probability may be approached from a slightly different angle through a consideration of certain principles related to permutations and combinations of specified sets of diverse units. From previous considerations it will be seen that the probability of two independent events occurring together, in cases like those discussed, is the product of their separate probabilities. Thus the probability of throwing three aces in three throws of one die or in one throw of three dice is $1/6 \times 1/6 \times 1/6 = 1/216$. Similarly in throwing two dice the chance of throwing an ace with the first and some other number with the second is $1/6 \times 5/6 = 5/36$. However, the total probability in a case of alternative odds, such as is illustrated in the chances of throwing an even number in one throw of a die, is the sum of the alternates $1/6 + 1/6 + 1/6 = 1/2$.

The permutations and combinations that may arise out of typical chance events may be simply illustrated by an elementary form of Mendel's law of heredity. When pollen grains, half of which carry a potency for blue flower-color (B) and half for white (W), fertilize seeds of similarly contrasting potencies the following permutations have equal chances of happening:

B fertilizes B

B fertilizes W

W fertilizes B

W fertilizes W

These four permutations, however, fall into only three combinations, since the second and third are practically alike. Hence arise the three classes and the $1:2:1$ frequency ratio of the binomial $(B + W)^2$; and similarly $(B + W)^4$ may be shown to fall into sixteen permutations, but only five combinations. In general, the frequencies ($2^s = n$) in the expansion of $2^s(1/2 + 1/2)^s$ are the permutations, and the classes $(s + 1)$ are the combinations.

It can be shown that the number of permutations (${}_nP_r$) comprising r diverse items each that can be drawn from a total of n diverse items is given by the formula:

$$ {}_nP_r = n(n-1)(n-2)\ldots(n-r+1), $$

and

$$ {}_nP_n = \lfloor\underline{n} $$

These permutations fall into a lesser number of combinations as expressed by the formula

$$_nC_r = {_nP_r} \div \underline{|r}.$$

The form which such combinations take for values of n and r from 0 up to 5 is shown in Table 13. The structure of the table obviously repeats the binomial $(1/2 + 1/2)^s$, and the sum of the combinations for any n, when r is ranged by units from 0 to n is, $\Sigma_nC_r = 2^n$; or $2^n - 1$, if $r = 0$ is not considered; while $\Sigma_nP_r = e\underline{|n}$ approximately, when n is large.

TABLE 13

The number of combinations ($_nC_r$) that can be made from n different items selected r at a time, where n and r range from 0 to 5, and $r = 0$ is taken as one combination.

	$n = 5$	4	3	2	1	0
r						
0	1	1	1	1	1	1
1	5	4	3	2	1	
2	10	6	3	1		
3	10	4	1			
4	5	1				
5	1					
	—	—	—	—	—	—
	$\Sigma_nC_r = 32$	16	8	4	2	1

Income distribution theory.—An application of the binomial distribution may be illustrated by the following problem. It is known that the curve of distribution of incomes in the United States approximates a logarithmic normal in its lower ranges, including about 85% of the incomes as arrayed from the lowest to the 15% highest. Supposing it is assumed that this distribution of income is a direct function of economic abilities, although, as a rule, biological capacities show somewhat less variability. On the basis of this assumption it might be asked what would be the highest earned income. The question assumes that unearned incomes in all except the highest brackets do not appreciably affect the curve of distribution, inasmuch as they might be taken as a constant factor of the given income. The largest incomes, however, may plausibly be assumed to represent a larger proportion of unearned income, that is, of income directly attributable to property (cf. Reinhardt and Davies, "Principles and Methods of Sociology," pp. 522–523). On the basis of the assumption just stated, the upper limit of the curve, considered as a point binomial on a logarithmic base, may be found as follows: The income-receiving population in 1918, on the basis of data

compiled by the National Bureau of Economic Research, was $n = 40,$-069,000. The position of the upper frequency (U) on a standard deviation scale is:

$$2^s = 40{,}069{,}000$$

hence $s \log 2 = \log 40{,}069{,}000$; $s = 25.256$; and $\sqrt{s} = 5.02553$.

But on a logarithmic normal curve fitted to the first and third quartiles ($Q_1 = \$833$, $Q_3 = \$1574$), the logarithmic standard deviation (L_s) is:

$$L_s = 0.5(\log 1574 - \log 833) \div 0.67449 = 0.20487$$

Hence on a scale centered at the logarithmic interquartile point, the position of U is located as

$$\log U = 0.5(\log 1574 + \log 833) + (5.02553 \times 0.20487)$$
$$= 4.08841, \quad \text{and} \quad U = \$12{,}258$$

But the upper limit thus determined is too low, by reason of the fact that the curve thus fitted has an average below the actual average, since many large incomes are excluded. The correction factor required may be found approximately as the ratio of the mean of the actual distribution ($AM = \$1543$), and the average of the abbreviated curve thus described. The average of the curve may be shown to be the geometric mean multiplied by the factor $\sqrt{a}$, where a is defined as:

$$a = \sigma_r^{L_s/\log e}$$

That is, a is the antilog of the expression $L_s^2 \div 0.4343$, or 1.24923. If the logarithmic interquartile point is taken as an approximation to the geometric mean with which theoretically it is identical, the arithmetic mean of the curve is \$1280 (i.e., $GM\sqrt{a}$), while the actual average income should be 1574, or 1574/1280 times larger. Hence the curve may be shifted on the logarithmic scale to bring the correct average, and as a result the upper limit becomes: $12{,}258 \times 1543 \div 1280 =$ 14,800, or practically \$15,000. That is, if the distribution of incomes in 1918 had followed throughout its entire range the curve which approximately fits the large majority of incomes, the largest income would have been about \$15,000.

General uses of binomial distributions.—The most common use, however, of the point binomial is in the calculation of standard and probable errors, and other probability functions. For example, it may be shown that if successive samples of n items each are drawn from an immensely large assortment, or statistical "universe," which is regularly distributed in the form $(1/2 + 1/2)^s$, the means of these

samples will be distributed in the form $(1/2 + 1/2)^{ns}$. The standard deviation of the "universe" expressed as a fraction of its range of s class intervals is $0.5\sqrt{s} \div s = 1 \div (2\sqrt{s})$; and of the means of the samples similarly expressed, $0.5\sqrt{ns} \div ns = 1 \div (2\sqrt{ns})$, the former being $\sqrt{n}$ times the latter. Hence under conditions of random sampling the dependability of a mean increases with the square root of the number of items. More precisely, the standard deviation of the means (σ_m) is the standard deviation of the assumed sample (σ) divided by the square root of the number of items in the sample, or $\sigma_m = \sigma \div \sqrt{n}$. Similarly, if a percentage sample (σ_s) of n items is taken, classified as p and q (as a straw vote of 80% for and 20% against), the standard deviation of $(p + q)^1$ in successive samples will vary about the mean, p or q, like σ_m above. The reliability of the percentage sample may therefore be estimated by measuring this deviation in terms of the whole range of class intervals as $\sigma_s = (pq/n)^{1/2}$, since $s = 1$. Thus a straw vote of 400 giving 80% for and 20% against a certain measure may be regarded as having a standard deviation of $\sigma_s = (pq/n)^{1/2} = (0.20 \times 0.80/400)^{1/2} = 0.02$, or 2 on a scale of 100, because this would presumably be the standard deviation of p or q, i.e., the mean of $(p + q)^1$, in successive samples of 400 drawn at random from a "universe" of votes actually divided in an 80 : 20 ratio. Hence, by reversing the logic, it might be assumed that an 80 : 20 distribution in a straw vote of 400 indicates a universe of uncertain distribution probably varying about 80% (or 20%) by $\sigma = 2\%$, with extreme limits at 0 and 100%.

Probable errors.—It is common practice to reduce the standard deviation in distributions such as $(1/2 + 1/2)^s$ to the quartile deviation, or probable error (0.67449σ), thus indicating the range about the measure in question within which 50% of the deviations will probably occur. Thus the standard deviations just referred to give the following probable errors, measuring the degree of unreliability.

Probable error of the mean $= 0.6745\sigma/\sqrt{n}$
Probable error of a percentage sample $= 0.6745(pq/n)^{1/2}$

This formula in a somewhat more complex form is applied later to frequency distributions. Other probable errors, analogous to the above, are:

Probable error of the standard deviation $= 0.6745\sigma/\sqrt{2n}$
Probable error of the quartile deviation $= 0.919\sigma/\sqrt{n}$
Probable error of a frequency $= 0.6745[f(n - F)/n]^{1/2}$
Probable error of the coefficient of correlation $= 0.6745(1 - r^2)/\sqrt{n}$
Probable error of the correlation ratio $= 0.6745(1 - \eta^2)/\eta$. Many other probable errors will be found in advanced text-books.

Misleading conclusions.—The application of the theory of probable errors is often a help in avoiding erroneous conclusions which are likely to be drawn from statistical data. Studies of both social and economic data, such as death rates and income distributions, are commonly drawn from large areas where the population is not homogeneous, but is made up of more or less clearly marked subgroups which have their own characteristics. Averages, or rates, in one city, state, or other unit compared with another, may be misleading because the population bases are not entirely comparable. A common example is that of death rates. It is possible that one country may have a lower death rate than another merely because of the composition of its population, rather than because of any essential differences in vitality or living conditions. In general, a frontier country to which immigrants are moving will have a relatively large proportion of its population in the active years of life, and death rates will appear low, while in an old country from which migration is taking place the ratès will appear high. Hence death rates when used for comparative purposes are commonly standardized; that is, a weighted average of specific age and sex groups is taken, in which the weights are proportional to the composition of a typical or standard population. The effect of variable groups is thus removed, and the average rates thus obtained become much more closely comparable from one country to another. In a similar way any rates reflecting economic or social conditions may be scrutinized for variable factors, and comparisons may be made in the light of such a study. Thus many erroneous conclusions may be avoided. This does not mean, of course, that broad studies of population are to be avoided, but merely that they should be made with due caution, and particularly that they should be supplemented by studies of the more homogeneous subgroups of which they may be composed.

The standard error (σ_R) of a rate is given by Professor Frank Alexander Ross (*American Journal of Sociology*, January, 1933) as follows:

$$\sigma_R = \sqrt{R(b - R) \div n}$$

where R is the rate per b; b is the denominator of the rate, as 100 or 1000; and n is the population. The variability of the difference between the two rates is

$$\sigma_{(R_1-R_2)} = \sqrt{\sigma_{R_1}{}^2 + \sigma_{R_2}{}^2}$$

and an accepted test of the significance of such a difference is

$$R_1 - R_2 > 3\sigma_{(R_1-R_2)}$$

If this inequality is reversed, it may be assumed that no real difference is indicated by the figures. These formulas are derived from the binomial equations already discussed. The application of such measures, including the Lexis ratio, to the study of subgroups is now becoming general in fields where large, diverse populations or data are considered.

Curve fitting.—Since, as has been seen, various types of distribution curves measure the relative probabilities of specific events occurring, the fitting of appropriate curves to data becomes a part of the analysis of such data in an attempt to discover the characteristics of the statistical field which they represent. The fitting of three curves will be described, the latter two of which may be considered to include the binomials when s is large.

The Poisson series.—After the binomial distributions, the Poisson series is perhaps the most important discontinuous curve of distribution. It is based on the assumption that the statistical probability varies from trial to trial in each repeated set of trials. As ordinarily computed, it forms a skewed curve of a rather high degree of skewness. Mathematically, it is easily fitted to data. Its principal use is in computing probability where the probability of occurrence is very small (cf. Fisher, R. A., "Statistical Methods for Research Workers," pp. 54–61).

The Poisson series may be compared with the point binomial by noting in both cases the ratios of any given term to the preceding term. In a symmetrical normal discontinuous curve, such as $(a + b)^s$, the coefficients of the expanded binomial are:

$$Y = 1 : \frac{s}{1} : s\frac{(s-1)}{1 \times 2} : \frac{s(s-1)(s-2)}{1 \times 2 \times 3} : \frac{s(s-1)(s-2)(s-3)}{1 \times 2 \times 3 \times 4}$$

When $s = 4$ this becomes

$$Y = 1 : \frac{4}{1} : \frac{4 \times 3}{\lfloor 2} : \frac{4 \times 3 \times 2}{\lfloor 3} : \frac{4 \times 3 \times 2 \times 1}{\lfloor 4}$$

The ratios (r) of each term to the preceding, taken in succession beginning with the second term, may be expressed as

$$r = \frac{s}{1} : \frac{s-1}{2} : \frac{s-2}{3} \cdots \frac{1}{s}$$

which, when $s = 4$, becomes

$$r = \frac{4}{1} : \frac{3}{2} : \frac{2}{3} : \frac{1}{4}$$

The Poisson series, however, takes a form based upon the arithmetic

mean (M) of the distribution. Assuming the x-scale to be 0, 1, 2, 3, etc., the distribution is as follows:

$$Y = 1 : M : \frac{M^2}{\underline{|2}} : \frac{M^3}{\underline{|3}}$$

carried out until the magnitude of Y becomes negligible. The area of this curve is e^M, where e is the constant 2.71828 ($\log e = 0.4343$). The ratio (r) of each term to the preceding, taken in succession beginning with the second term, is:

$$r = \frac{M}{1} : \frac{M}{2} : \frac{M}{3} \ldots \text{etc.}$$

It follows from the nature of the Poisson series that it may be readily fitted to a given distribution by assuming the x-scale 0, 1, 2, 3, etc., and calculating the arithmetic mean on the basis of such an assumed scale. The curve may then be calculated by the formulas just given, as illustrated in Example 85. The area of the computed curve may be adjusted to the area of the data by computing the totals in each case and multiplying each frequency in the computed curve by the ratio of the data total (Σf) to the computed curve total (ΣY).

It should be noted, however, that the Poisson series is seldom adaptable to social and economic data, although, as previously suggested, it has important uses in calculating certain classes of probabilities.

Example 85.—Poisson series fitted to an assumed distribution. For the given X-scale there is substituted the x-scale 0; 1; 2; 3; etc. The arithmetic mean (M) of the distribution on this scale is calculated. The Poisson series (Y) is taken as $Y = 1 : M : M^2/\underline{|2}; M^3/\underline{|3} . : .$ carried out to a point where the magnitude of Y becomes negligible. This calculation is readily made by noting that the successive ratios (R) of each term to the preceding, beginning with the second, are: $R = M/1 : M/2 : M/3 :$ etc. The sum of the Y's should check approximately as $\Sigma Y = 2.71828^M$ (or antilog $0.4343M$), and this area is adjusted to the area of the original frequencies by multiplying each frequency by the ratio $\Sigma f/\Sigma Y$.

Data, assumed distribution:

														Σ
f:	3	10	14	21	22	18	11	5	3	2	1	0	1	0	111
X:	2	4	6	8	10	12	14	16	18	20	22	24	26	28	
fX:	6	40	84	168	220	216	154	80	54	40	22	0	26	0	1110
x:	0	1	2	3	4	5	6	7	8	9	10	11	12	13	
fx:	0	10	28	63	88	90	66	35	24	18	10	0	12	0	444

Arithmetic mean on X-scale, 10; on x-scale, $4 = M$.

Poisson series, or distribution (Y):

Y:	1	4	8	10.67	10.67	8.53	5.69	3.25	1.63	0.72	0.29	0.11	0.04	0.01	54.61

Ratio $\Sigma f/\Sigma Y = 111/54.61 = 2.0326$; $Y \times \Sigma f/\Sigma Y = Y$ adjusted (Y_a).

Y_a:	2.03	8.13	16.26	21.69	21.69	17.34	11.57	6.61	3.31	1.46	0.59	0.22	0.08	0.02	111.00

The normal curve.—This curve is mathematically described by the equation: $Y = \frac{N}{\sqrt{2\pi}} e^{-x^2/2}$ where x is expressed in units of the standard deviation from the mean, and N is the total area. The integral, or area (A) of the curve from $x = 0$ to any x, is expressed by the converging series

$$A = \frac{N}{\sqrt{2\pi}} \int e^{-\frac{x^2}{2}} dx = \frac{Nx}{\sqrt{2\pi}} \left(1 - \frac{x^2}{2 \cdot 3 \cdot \underline{|1}} + \frac{x^4}{2^2 \cdot 5 \cdot \underline{|2}} - \frac{x^6}{2^3 \cdot 7 \cdot \underline{|3}} + \cdots \right)$$

But in practical work, tables are generally used to obtain the ordinates and integrals.

For purposes of calculation the equation of the normal curve* may be rewritten

$$Y = \frac{0.3989\, ni}{\sigma \text{ antilog } [(0.21715/\sigma^2)\,(X - M)^2]}$$

where Y is the height of the curve at any point, X, on the base; n is the sum of the frequencies; i is the class interval; σ is the standard deviation; and M the arithmetic mean. The formula may be separated into two factors: Area $= ni/\sigma$ (the total area of the distribution in σ and f units), and $z = 0.3989 \div [\text{antilog } (0.21715/\sigma^2)(X - M)^2]$. The value of z may be read from a table of the normal curve of unit area as the ordinate of $x/\sigma = (X - M)/\sigma$. Thus the ordinate of the curve at any selected point X on the base is $Y = z(ni/\sigma)$. The normal frequencies may be computed by differencing the areas as read from a table of the probability integral for the class limits as X. The process is illustrated in Example 86. For convenience the ordinates are computed at the class limits, but the same method may be applied to the class marks. An additional class beyond the frequencies is affixed at each extreme; in some problems two or three such classes may be required in order to reduce the end frequencies to negligible size.

Example 86.—Normal ordinates (Y) and normal frequencies (F) fitted to data by the formulas: $Y = z(ni/\sigma)$; $F = n\Delta$ areas; read table of ordinates and integral

* The normal curve may be used to express the probable occurrence of a deviation expressed in units of the standard deviation. Thus the probability that a deviation greater than 1 will occur is 31.7%. This figure may be obtained by noting that in the table of the normal curve of error the area at $x = 1.00$ is 0.3413, or 0.6826 if both sides of the curve are considered. The remainder of the curve of unit area is therefore $1.00 - 0.6826$, or 0.3174. The odds against such a deviation occurring are 0.6826/0.3174 or 2.15 to 1.00. In the same way, other probabilities may be computed. A convenient table of these probabilities is given in "Medical Biometry and Statistics," by Pearl, and this table is reprinted in "Mathematical Tables," by Hodgman.

of the normal curve at $x/\sigma = (X - M)/\sigma$, where L_2 is selected as X. The arithmetic mean (M) was previously computed as 5.6, and σ as 1.8. The area is differenced from -0.5 to 0.5 to obtain Δ; the two extreme F's represent sums of successive diminishing frequencies. If the distribution is irregular it may be better to take the standard deviation (σ) as $QD \div 0.67449$, obtaining the quartiles by the more exact formula given in the chapter on dispersion.

Data L_1	$L_2(X)$	f	x $X-M$	x/σ $(X-M)/\sigma$	Cf. table z	$z(ni/\sigma)$ Y at L_2	Cf. table Area	Differences Δ_1	$n\Delta_1$ F
0 —	2	0	−3.6	−2.00	0.054	0.60	−0 477	(0.023)	0.23
2 —	4	2	−1.6	−0.89	0.268	2.98	−0.313	0.164	1.64
4 —	6	4	0.4	0.22	0.389	4.32	0.087	0.400	4 00
6 —	8	3	2.4	1.33	0.165	1.83	0.408	0.321	3.21
8 —	10	1	4.4	2.44	0.020	0.22	0 493	0.085	0.85
	(12)	0	.....	......	.....	....	....	(0.007)	0.07
		$n=10$							

The logarithmic normal curve.—Most distributions which approach the normal form have a definite skewness. As a rule, these distributions may be fitted by the logarithmic normal curve, either directly or as adapted to the given degree of skewness.

To test whether the logarithmic normal curve is adapted to a given distribution, we may compute the quartiles and find the logarithmic coefficient of skewness:

$$Sk_1 = (\log Q_1 + \log Q_3 - 2 \log Q_2) \div (\log Q_3 - \log Q_1)$$

If this coefficient is less than 0.15, or perhaps even 0.20, the distribution may tentatively be regarded as of the logarithmic normal type. If, however, it is not taken as inherently logarithmic, then the fitted curve may be adjusted to the given degree of skewness by shifting the data on the X-scale to a position where it will be logarithmic. The correction (c) to be added to all the X-magnitudes (the class marks, class limits, etc.) is found by means of the quartiles, as follows:

$$c = (Q_2^2 - Q_1Q_3)/(Q_1 + Q_3 - 2Q_2)$$

After the curve has been computed (cf. Example 87), the correction (c) is subtracted from the class marks, the mode, and other X-magnitudes. When this correction is employed, the fitted curve has the same degree of skewness as the data, as measured above.

For the fitting of the curve, two measures are required, namely, the logarithm of the geometric mean ($\log G$) and the logarithms of the standard deviation ratio, i.e., the logarithmic standard deviation ($\log \sigma_r$). Unless the data are very regular, these may be more satisfactorily computed from the quartiles, thus avoiding the disturbing effects

of irregular extreme items. If the correction is not used, the formulas mentioned in an earlier chapter are employed; namely,

$$\log G = (\log Q_1 + \log Q_3 + 1.2554 \log Q_2)/3.2554$$

which is an average of the logarithmic quartiles weighted in proportion to their normal ordinates, and

$$\log \sigma_r = 0.7413 (\log Q_3 - \log Q_1)$$

which is half the logarithmic interquartile range divided by 0.67449.

If, however, the correction (c) is computed and used, then $\log G$ is merely $\log Q_2$ as corrected, and $\log \sigma_r$ is found as before.

If an estimate of the mode is desired, it may be found without actually fitting the curve, as may also other measures such as the standard deviation. For this purpose the following formulas may be employed.

As a preliminary step find the constant a by the formula,

$$a = \text{antilog}\ \overline{(\log \sigma_r)^2/0.4343}$$

Then $$Mo = G/a$$

(If the geometric mean, G, had been found from quartiles corrected by adding c, then c obviously should be subtracted from the mode thus found.)

Also, the mean of the smoothed curve is

$$AM = Ga^{1/2}$$

(Subtract c if necessary, as just suggested for the mode.)
And the standard deviation of the smoothed curve is

$$\sigma = G[a(a-1)]^{1/2}$$

Moments (v) of the smoothed curve from an origin of zero may be obtained by the formula

$$v_m = G^m a^{m^2/2}$$

The Betas ($\beta_1 = \mu_3^2/\mu_2^3$ and $\beta_2 = \mu_4/\mu_2^2$) as used in the Pearsonian system of frequency curves are functions of the moments (μ) about the mean, and may be calculated for the fitted curve by the formulas:

$$\beta_1 = a^3 + 3a^2 - 4$$

$$\beta_2 = a^4 + 2a^3 + 3a^2 - 3$$

The logarithmic normal curve may be fitted to data, whether corrected

for skewness or not, by the use of the following equation (unit area; notation as previously used and $L_s = \log \sigma_r$)

$$Y = G \div X(2\pi)^{\frac{1}{2}} e^{\frac{1}{2}\left[\left(\log \frac{X}{G}\right) \div L_s\right]^2}$$

or, as adjusted for purposes of calculation (area as given),

$$Y = \frac{0.17326\, ni/L_s}{X \text{ antilog } [(0.21715/L_s^2)(\log X - \log G)^2]}$$

Example 87.—Logarithmic normal curve, expressed both as ordinates and frequencies, fitted to the distribution of magnitudes (m) and frequencies (f), as follows:

$$m = 22,\quad 26,\quad 30,\quad 34$$
$$f = 2,\quad 4,\quad 3,\quad 1$$

The quartiles have been computed on the basis of a smoothing process as follows:

$$Q_1 = 24.570; \quad Q_2 = 27.019; \quad Q_3 = 29.675$$

The correction (c) to be added to each X-magnitude (class limits, class marks, quartiles, etc.) in order to shift the distribution to a point where it becomes logarithmic as measured by the quartiles, is as follows:

$$c = (Q_2^2 - Q_1Q_3)/(Q_1 + Q_3 - 2Q_2) = (730.0158 - 729.1265)/0.2079 = 4.278$$

G is $Q_2 + c = 31.297$, and $\log \sigma_r$ is obtained by the following formula:

$$\log \sigma_r = [\log (Q_3 + c) - \log (Q_1 + c)] \times 0.7413 = 0.7413(1.5309 - 1.4601)$$
$$= 0.05246$$

The ordinates of the logarithmic normal curve are fitted for convenience at the upper limits of the classes, though other ordinates may be found by the same process. The upper limits are advanced on the X-scale by adding $c = 4.278$. The process is as follows:

Find $x/\sigma = (\log X - \log G)/\log \sigma_r$; read ordinate ($z$) and area from table of normal curve of unit area; take $Y = (0.4343\, ni/\log \sigma_r)z/X$; Δ_1 as first differences of area from -0.5 to 0.5; and F as Δ_1 times $n = 10$; taking $\log G = 1.4955$, and $\log \sigma_r = 0.05246$. The two extreme F's contain small residuals belonging to more extreme frequencies. The last column shows the deviations of the data from the normal in units of the probable error of sampling. The calculations were carried to more decimal places than are here indicated.

L_2	f	X	Log X	x/σ	z	Y	Area	Δ_1	F	d/PE_s
16	0	20.278	1.3070	−3.5926	0 0006	0.0101	−0.4998	0.0002	0.002	0.07
20	0	24.278	1.3852	−2.1024	0.0438	0.5970	−0.4822	0.0176	0.176	0.63
24	2	28.278	1.4514	−0.8397	0.2804	3.2837	−0.2995	0.1828	1.828	0.21
28	4	32.278	1.5089	0.2556	0.3861	3.9611	0.1009	0.4003	4.003	0.00
32	3	36.278	1.5596	1.2226	0.1889	1.7245	0.3893	0.2884	2.884	0.12
36	1	40.278	1.6051	2.0886	0.0450	0.3703	0.4816	0.0924	0.924	0.12
40	0	44.278	1.6462	2.8725	0.0064	0.0482	0.4980	0.0163	0.163	0.60
44	0	48.278	1.6837	3.5884	0.0006	0.0043	0.4998	0.0019	0.020	0.21

To find the Mode (Mo) and the ordinate at the mode (Y_{Mo}):
Let $a = \text{antilog}\,[(\log \sigma_r)^2/0.4343]$
Then $Mo = G/a = 31.297/1.0147 = 30.844$
and $Mo - c = 26.566$ on original X-scale.
$Y_{Mo} = (0.17326\, ni/\log \sigma_r)\, a^{\frac{1}{2}}/G$
$= 132.108 \times 1.00732 \div 31.2968 = 4.2520.$

or, when z is read from a table of ordinates of the normal curve of unit area at $x/\sigma = (\log X - \log G)/L_s$, the following equation is preferable:

$$Y = (0.4343\, ni/L_s)z/X$$

An example of such curve fitting may be found in the *Journal of the American Statistical Association*, December, 1929, p. 362. It is reproduced, by permission, in Example 87. The normal frequencies (F) are computed by the use of the table of the integral of the normal curve, as indicated.

The probable error of sampling.—The standard error (σ_s) of sampling as applied to frequency distributions is a measure of the variability of the frequencies of a distribution from the frequencies of a curve fitted to them. The so-called probable error of sampling (PE_s) is 0.6745 times the standard error thus found. It is computed as follows:

$$PE_s = 0.6745 \times [F(n - F)/n]^{1/2}$$

The given frequencies (f) may be substituted for the normal frequencies (F) in this formula without significant changes in the conclusions drawn. The probable error of sampling tests the variability of the data from the smooth curve by comparing it with that which is theoretically likely to occur. Hence the deviations of the data (f) from the computed curve (F), taken without regard to signs (i.e., $d = |f - F|$), divided by the probable error of sampling, should, on the basis of chance, range from zero to perhaps three or four. If the ratios (d/PE_s) are consistently less than unity a good fit is indicated. For more refined methods of testing goodness of fit, see the Chi-Square test and similar methods explained by R. A. Fisher in "Statistical Methods for Research Workers." The calculation of the measure of "goodness of fit" for the curve calculated in Example 87 is given in Example 88.

Example 88.—The measure of "goodness of fit," for the curve fitted in Example 87. The standard error of sampling is $\sigma_s = [F(n - F)/n]^{1/2}$, and the probable error of sampling (PE_s) is $0.6745\sigma_s$.

f	n	F	$n-F$	$F(n-F)$	$F(n-F)/n$	σ_s	PE_s	$d=\|f-F\|$	d/PE_s
0	10	0.002	9.998	0.020	0.0020	0.045	0.030	0.002	0.07
0	10	0.176	9.824	1.729	0.1729	0.416	0.281	0.176	0.63
2	10	1.828	8.172	14.938	1.4938	1.222	0.824	0.172	0.21
4	10	4.003	5.997	24.006	2.4006	1.549	1.045	0.003	0.00
3	10	2.884	7.116	20.523	2.0523	1.433	0.967	0.116	0.12
1	10	0.924	9.076	8.386	0.8386	0.916	0.618	0.076	0.12
0	10	0.163	9.837	1.603	0.1603	0.400	0.270	0.163	0.60
0	10	0.020	9.980	0.200	0.0200	0.141	0.095	0.020	0.21

EXERCISES

1. If 5 coins are thrown simultaneously, what are the odds of throwing all heads? All tails? If 10 coins are thrown? *Ans.* 1 in 32; 1 in 1024.

2. If 4 dice are thrown simultaneously, what are the odds of throwing all aces? All sixes? *Ans.* 1 in 1296.

3. If 5 coins are thrown simultaneously what are the odds of throwing either all heads or all tails? Either two heads and three tails or three heads and two tails? *Ans.* 1 in 16; 20 in 32.

4. If a random straw vote of 10,000 gives a 50 : 50 ratio, what possibilities may be assumed regarding the field? A 90 : 10 ratio?

Ans. $50\% \pm \overline{\sigma = 0.5\%}$; $90\% \pm \overline{\sigma = 0.3\%}$.

5. Fit a normal curve to the point binomials $(\frac{1}{2} + \frac{1}{2})^4$, $\sigma = 1$, and $(\frac{1}{2} + \frac{1}{2})^6$, $\sigma = 1.2247$, determining ordinates at the class limits and class marks ($m = 0, 1, 2 \ldots s$) and normal frequencies at the class marks.

6. Check the following curve fitting on the assumption that curves are of the logarithmic normal type. Plot data and normal.

(A) Price relatives, 1924/1913.

m	20	40	60	80	100	120	140	160	180	200	220
$f\%$	0.3	0	1.0	2.6	7.2	11.6	19.4	17.3	14.7	9.8	8.8
m	240	260	280	300	320	340	360	380	400	420	440
$f\%$	2.3	2.9	0.5	1 0	0	0	0	0	0.3	0	0.3

$$\log Q_1 = 2.12234; \quad \log Q_2 = 2.20088; \quad \log Q_3 = 2.28230$$

Logarithmic normal computed at successive σ_r points (X), centering at $X = 159.16$.

X	70.16	80.43	92.19	105.67	121.13	138.85	159.16
Y	0.46	1.60	4.29	8.98	14.63	18.57	18.36
X	182.44	209.13	239.72	274.79	314.98	361.06	Mo = 147.73
Y	14.14	8.48	3.96	1.44	0.41	0.09	19.06

(B) Price relatives, 1924/1913.

m	50	70	90	110	130	150	170	190	210
$f\%$	0.7	2.7	6.0	11.0	16.0	17.5	18.7	11.0	8.0
m	230	250	270	290	310	330	350	370	390
$f\%$	5.5	1.2	0.5	0.5	0.3	0.2	0	0	0.2

$$\log Q_1 = 2.10053; \quad \log Q_2 = 2.19138; \quad \log Q_3 = 2.26421$$

Logarithmic normal computed at successive σ_r points (X), centering at $X = 153.24$.

X	66.28	76 21	87.64	100.78	115.89	133.26	153.24
Y	0.48	1.65	4.41	9.20	14.95	18.91	18.64
X	176.21	202.63	233.01	267.94	308.11	354.31	
Y	14.30	8.55	3.98	1.44	0.41	0 09	

7, Check the following logarithmic normal distribution (F) fitted to a distribution of 241 towns in England and Wales classified by case rates (per 1000 living) of infectious diseases. Plot data and normal.

m..........	1	3	5	7	9	11	13
$f\%$.........	2.1	16.2	28.6	17.0	12.0	9.1	6.6
F..........	2.0	17.0	24.0	19.8	13.2	8.4	5.6
m..........	15	17	19	21	23	25	27
$f\%$.........	2.9	2.1	1.2	1.7	0	0	0.4
F..........	3.5	2.1	1.5	0.9	0.6	0.4	0.3

8. Check the following curve fitting and accompanying calculations:

Variability of length of business cycle in the United States, 1796–1923

(Mitchell, "Business Cycles," p. 399) Analysis by logarithmic normal curve based on quartiles interpolated by sloped frequencies.

(1) Duration Years (m)	(2) No. of Cycles f	(3) Normal F	(4) Deviations (d) (2)–(3)	(5) Fit d/PE
1	1	0.68	0.32	0.58
2	4	5.21	−1.21	−0.86
3	10	8.15	1.85	1.11
4	5	6.96	−1.96	−1.25
5	6	4.65	1.35	1.00
6	4	2.77	1.23	1.15
7	1	1.58	−0 58	−0.70
8	0	0.88	−0.88	−1.41
9	1	0.49	0.51	1.09
10	0	0.27	−0.27	−0.77

$Q_1 = 2.805$	$A = 4.155\ (4.031)^*$	$\sigma = 1.975\ (1\ 686)^*$
$Q_2 = 3.672$	$G = 3.752\ (3.676)^*$	$\beta_1 = 2\ 352\ (0.588)^*$
$Q_3 = 5.157$	$Mo = 3.061\ (3.0455)^*$	$\beta_2 = 7.453\ (3.578)^*$

* Usual method; substituting the fitted curve materially increases the higher moments. Sloping the frequencies does not change A, and decreases σ only from 1.686 to 1.669. The correction for sloped frequencies is unimportant here, since the distribution is irregular and the number of classes adequate. The betas by the usual method, without Sheppard's correction, are 0.538 and 3.544 respectively.

APPENDIX

MATHEMATICAL NOTES

Statistical notation.—The student of business statistics in the social sciences will not find the higher branches of mathematics necessary, but he should be familiar with algebraic notation, logarithms, etc. It should be recalled that in a series of factors and addends, plus and minus signs separate terms; for example: $5 + 2 \times 3 - 1$ is grouped $5 + (2 \times 3) - 1 = 10$. The expressed signs, $\times$ and $\div$, are read in succession, as $10 \div 5 \times 2 = 4$; but the slant line expressing division governs to the end of the term, as, $10/5 \times 2 = 1$. Also, in multiplication and division, like signs give plus, and unlike, minus. When plus and minus numbers are added without regard to the signs, the result is called the absolute or arithmetic sum rather than the algebraic sum. Such a sum is designated by vertical bars or single quotation signs. Thus, the summation of the deviations: -2, $+5$, -3; is written, $\Sigma \mid d \mid$, or $\Sigma' d' = 10$, though the algebraic sum is zero; that is, $\Sigma d = 0$.

It should be noted that percentages and usually inaex numbers carry two decimal places (hundredths), understood. For example, 80 (%) × 25 (%) = 20 (%), whether the per cent signs are written or merely implied. To avoid errors in more extended computations, such figures should be written with the decimal expressed, as $0.80 \times 0.25 = 0.20$.

The reciprocal of a number is the number itself divided into unity, thus the reciprocal of 5 is $\frac{1}{5}$ or 0.2, and the reciprocal of 100 is $\frac{1}{100}$ or 0.01. If the decimal reciprocal of a large number is to be obtained, difficulty may be encountered in determining the location of the decimal point. The following rule will be of use as a guide. Place the first significant figure of the calculated reciprocal under the units place of the number, and put the decimal point of the reciprocal as many places to the left of the decimal point of the number as there are places preceding the decimal in the number. For example, suppose the reciprocal of 251,867.432 is desired. The reciprocal as calculated to five significant figures is 39,703. The following form will place the decimal (except in the case of 10^n):

	251867.	432
0.	000003	9703

Thus the decimal reciprocal is 0.0000039703. In index numbers the reciprocal of the index (percentage) is often required, as the reciprocal of 20, 80, and 150—which may be written as $1/0.20 = 5$, $1/0.80 = 1.25$, and $1/1.50 = 0.67$, or as index numbers 500, 125, and 67, respectively.

The use of the summation sign.—A word of caution may be in order regarding the use of the summation sign in algebraic transformations. It must be remembered that Σ is not a factor but is a symbol meaning the sum of n terms using successive values of the related variables in the expression following the symbol. The summa-

tion is applicable only to the next plus or minus sign except as modified by a parenthesis. Thus, $\Sigma xy + c$ implies a summation of the terms ($xy + x'y' + x''y''$, etc.) to which total is added c, and is not the same as $\Sigma(xy + c)$. It follows, then, that if a constant, as b, occurs in a summated term, it may be written before Σ, that is, outside the indicated parenthesis. Thus $\Sigma bx = b\Sigma x$. Also the summation of a constant, as M, is n times the constant. Thus, $\Sigma M = nM$, $\Sigma(xy + c) = \Sigma xy + nc$, and $\Sigma(\Sigma x^2) = n\Sigma x^2$. In this connection it should be noted that though x, xy, x^2, and similar terms are variables, their summations as Σx, Σxy, and Σx^2, are constants. A careful distinction should always be made between variables and constants. Variables preceded by Σ cannot be canceled as factors, and powers or roots cannot be applied to these variables. Thus $\Sigma xy + \Sigma x^2$ cannot be simplified, and the square of Σx^2 is $(\Sigma x^2)^2$, not Σx^4. In summing fractions a constant denominator is unchanged; as, $\Sigma(x/\sigma) = \Sigma x/\sigma = (1/\sigma)\Sigma x$, and $\Sigma(xy/\Sigma x^2) = \Sigma xy/\Sigma x^2$.

Scales.—Various scales are used in calculation and in graphic representation, the two most important of which are the arithmetic (ordinary) and geometric (ratio or logarithmic). An arithmetic scale or progression proceeds by equal intervals, as 10, 20, 30, etc.; a geometric scale or progression proceeds by equal ratios; as, 1, 2, 4, 8, etc. The relation between the arithmetic and geometric scales is expressed by logarithms, the logs of a geometric series forming an arithmetic series. A table of logarithms gives numbers at regular intervals from 1 to 10, and the corresponding logs, from 0 to 1. The common logarithm is in theory the exponent of 10 required to equate the given number; hence, a number between 10 and 100 will have a log between 1 and 2, etc. A brief discussion of logarithms is given later in connection with a table of logarithms (cf. p. 331).

A reciprocal scale has many uses, as in distributing overhead costs in accounting, and in plotting Pearl–Reed curves. Ruled paper thus scaled may be obtained of dealers, or paper may be ruled by designating equal intervals, thus:

Intervals...........	0	0.2	0 4	0.6	0.8	1.0	1.2 etc.
Designation........	∞	5.00	2.50	1.67	1.25	1.00	0.83 etc.

The scale may be varied by graduating the intervals narrower or wider, to suit the purpose at hand. The data are plotted according to the designations; thus in effect the reciprocals are plotted.

Calculation of roots.—If roots are required to a greater degree of accuracy than can be obtained by the use of logarithms or other tables, the following method, which is adapted to a calculating machine, may prove convenient:

Given the number, n^x, of which the xth root is to be found.

1. Make an estimate (g) of the required root
2. Revise this estimated root (g) by the formula,

$$g_2 = \frac{n^x - g^x}{xg^{x-1}} + g$$

3. Repeatedly revise the estimate by the same formula until the change resulting from revision is less than the allowable margin of error.

Illustration: Find the fifth root of 4,084,101.

Estimate $g = 20$, as the root; then,

$$g_2 = \frac{4{,}084{,}101 - 3{,}200{,}000}{5 \times 160{,}000} + 20 = 21.1051$$

$$g_3 = \frac{4{,}084{,}101 - 4{,}187{,}328.9}{5 \times 198{,}403.65} + 21.1051 = 21.0010$$

Further revisions may be made if the root is required to a greater degree of accuracy. The formula is readily proved by means of the calculus derivative.

The constant of "organic growth"—The "growth" constant, $e = 2.718+$, is found in many complex mathematical formulas. Its significance and derivation may best be illustrated in connection with compound interest, as follows: Let us consider the case of one dollar put at interest for an interval of time long enough so that it earns 100% at simple interest; that is, the amount at the end of the time is double the investment. This is equivalent to saying that the interest is compounded only once, namely, at the end of the interval. If, however, the interest is compounded at the end of each half interval, the rate of interest is 50% for the half interval, and the final amount may be expressed as $(1 + \frac{1}{2})^2$. Similarly if the interest is compounded at the end of each quarter interval, the amount becomes $(1 + \frac{1}{4})^4$. If the number of times that the investment is compounded is further increased, the amount is further increased, up to a certain limit, thus:

Interest compounded:

once (at end of interval)	$(1 + 1)^1 = 2$
twice (at middle and end)	$(1 + \frac{1}{2})^2 = 2.25$
five times	$(1 + \frac{1}{5})^5 = 2.49$
ten times	$(1 + \frac{1}{10})^{10} = 2.59$
one hundred times	$(1 + \frac{1}{100})^{100} = 2.70$
one thousand times	$(1 + \frac{1}{1000})^{1000} = 2.72-$

The values thus found may be expressed in general by a binomial expansion, as

$$(1 + 1/x)^x = 1 + x\cdot\frac{1}{x} + \frac{x(x-1)}{\underline{|2}}\cdot\frac{1}{x^2} + \frac{x(x-1)(x-2)}{\underline{|3}}\cdot\frac{1}{x^3}, \text{ etc.}$$

where the factorial, $\underline{|2} = 1\cdot 2$; $\underline{|3} = 1\cdot 2\cdot 3$; $\underline{|4} = 1\cdot 2\cdot 3\cdot 4$, etc. If, now, x approaches infinity, x, $x - 1$, $x - 2$, etc., are practically equal, and the equation becomes:

$$e = 1 + 1 + \tfrac{1}{2} + \tfrac{1}{6} + \tfrac{1}{24}, \text{ etc.} = 2.71828+$$

Mathematical proofs for some of the more important theorems in statistics pertinent to the preceding chapters, together with certain formulas and their uses, are suggested in the following pages. They are classified according to the headings of the chapters to which they are most closely related.

Averages

The deviations from the arithmetic mean (M) balance; i.e., $\Sigma d = 0$. Proof: $\Sigma d = \Sigma(X - \Sigma X/n) = \Sigma X - n\Sigma X/n = 0$. (The summation of a constant, as $\Sigma X/n$, is n times the constant.)

The absolute deviations ($\Sigma' d'$) from the median are a minimum. This may be demonstrated by plotting the items and the median point as the origin on a horizontal scale, expressing the deviations as lines parallel to the scale, and observing the effect of shifting the origin past one of the plotted items (see Chart 7).

If a constant (C) is added to each of a series of items (m), the arithmetic mean (M) is increased by that constant

$$\Sigma(m + C)/n = (\Sigma m + nC)/n = M + C$$

The sum of the squares of the deviations about the arithmetic mean (M) is a minimum.

$$\Sigma d^2 = \Sigma(X - \Sigma X/n)^2 = \Sigma X^2 - 2\Sigma X\ \Sigma X/n + n(\Sigma X)^2/n^2 = \Sigma X^2 - (\Sigma X)^2/n$$

Change M by adding C:

$$\Sigma d^2 = \Sigma(X - \Sigma X/n - C)^2 = \Sigma X^2 + (\Sigma X)^2/n + nC^2 - 2\,(\Sigma X)^2/n - 2C\Sigma X + 2C\Sigma X = \Sigma X^2 - (\Sigma X)^2/n + nC^2$$

which is larger than before.

A constant factor (C) included in a series of positive weights (w) applied to the positive variable numbers (m) will not change the resulting arithmetic, geometric, or harmonic means, since C carried into the product is canceled by division. Hence the ratio of the weights, rather than their differences, is significant. Also, the constant (C) added to each weight causes the ratio of one weight to any other to approach unity, thus, as C increases, progressively neutralizing the effect of the weights. This is evident if the average (weights as $w + C$) is written:

$$\frac{\Sigma mw + C\Sigma m}{\Sigma w + Cn}$$

which is a weighted mean of $\Sigma mw/\Sigma w$ and $C(\Sigma m/n)$, where n is the number of m's, and the weights are Σw (the original weights) and nC, respectively. The principle holds for the average of m, $\log m$, and $1/m$.

The quadratic mean (Q) of a variables series (m) is greater than the arithmetic mean (M).

Write $d = m - M$, then $m = M + d$. And $\Sigma d/n = 0$.

$$Q^2 = \Sigma m^2/n = \Sigma(M + d)^2/n$$
$$= M^2 + 2M\Sigma d/n + \Sigma d^2/n = M^2 + \Sigma d^2/n$$
$$Q^2 > M^2 \quad \text{and} \quad Q > M$$

For two unequal positive numbers, the arithmetic (M), geometric (G), and harmonic (H) means give the inequalities:

$$M > G > H$$

Write the two numbers a and b as

$$a = M - d, \quad \text{and} \quad b = M + d$$

then $$(M - d)\,(M + d) = M^2 - d^2 = ab = G^2$$

hence $$M > G$$

But $MH = G^2$, since

$$[(a + b)/2] \times [2/(1/a + 1/b)] = ab = G^2$$

hence $$M/G = G/H$$

and $$M > G > H$$

Multiplying each of the items of a series by the constant C, multiplies their arithmetic mean (M), geometric mean (G), and harmonic mean (H) by C.

$$\Sigma Cm/n = C\Sigma m/n = C + M$$

$$(Cm_1 \cdot Cm_2 \cdot \ldots \cdot Cm_n)^{1/n} = C(m_1 \cdot m_2 \cdot \ldots \cdot m_n)^{1/n} = C \cdot G$$

$$n/\Sigma(1/Cm) = Cn/\Sigma(1/m) = C \cdot H$$

The arithmetic mean (M) of n variable items $(m_1 + m_2 \ldots + m_n)$ is greater than the geometric mean (G), which is greater than the harmonic mean (H).

Assume a variable series $(m_1 + m_2 \ldots + m_n)$ having a geometric mean of unity. Plot these items on the line $Y_1 = m$ (diagonal line of Chart 41). Plot also $1 + \log_e m$ and $2 - 1/m$ on each m ordinate, respectively. These points fall on the curves $Y_2 = 1 + \log_e m$, the slope of which is $1/m$; and $Y_3 = 2 - 1/m$, the slope of which is $1/m^2$. Then the curves Y_1, Y_2, and Y_3 are tangent at $m = 1$, since each Y and its slope is unity at that point; and as indicated by the slopes, at all other points Y_1 is above Y_2, and Y_2 is above Y_3. Hence, except when all m's are equal

$$\Sigma Y_1 > \Sigma Y_2 > \Sigma Y_3$$

Substituting the equations of the curves and dividing by n,

$$M > 1 + \log_e G > 2 - 1/H$$

Since $\qquad \log_e G = 0; \quad M > 1$

and $\qquad -1 > -1/H; \quad 1 > H$

since $\qquad G = 1, \quad M > G > H$

The proof may be considered general since, if the m's are each multiplied by a constant, the three averages are also multiplied by the same constant.

The average of positive variables (m) and the average of their reciprocals ($1/m$), using identical weights with m and $1/m$, respectively, will give a product greater than unity. For the geometric mean of m and $1/m$ is unity; that is, (assuming subscript of m with w)

$$(m_1^w m_2^w \ldots m_n^w)^{1/n} \times (1/m_1^w \cdot 1/m_2^w \ldots 1/m_n^w)^{1/n} = 1$$

But the arithmetic mean is greater than the geometric mean, hence,

$$(\Sigma mw/\Sigma w) \times (\Sigma \overline{1 \div m}\, w/\Sigma w) > 1$$

For this reason the average of price relatives (m) forward and backward, using identical weights with m and $1/m$, is said to have an upward bias.

The differences between the arithmetic (M), geometric (G), and harmonic (H) means of a variable series of numbers (m) are lessened if a constant (C) is added to the numbers.

Assume the conditions represented in Chart 41, and add an infinitesimal (C) to each m. Then, including the summations of the slopes,

$$M + C > 1 + C\Sigma(1/m)/n > 2 - 1/H + C\Sigma(1/m^2)/n$$

but $\qquad 1 < \Sigma(1/m)/n < \Sigma(1/m^2)/n$

since $\Sigma(1/m)/n = 1/H > 1$, and $\Sigma(1/m^2)/n = Q^2$ of $1/m >$

$$M^2 \text{ of } 1/m > 1/H > 1 = G \text{ of } 1/m$$

Hence the inequalities $M > G > H$ are lessened.

The proof is readily extended to series of any value of G.

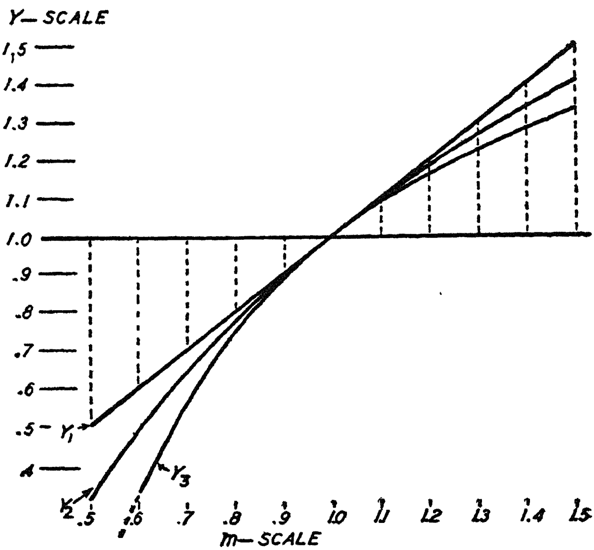


CHART 41

Illustration of the proof that the arithmetic (M), geometric (G), and harmonic (H) means of a positive variable (m) are unequal in the sequence $M > G > H$. The curves $Y_1 = m$, $Y_2 = 1 + \log_e m$, and $Y_3 = 2 - 1/x$, are plotted against the horizontal m scale, where the geometric mean of the given m's is unity. The slopes of the curves, which are tangent at $m = 1$, are 1, $1/x$, and $1/x^2$, respectively. From the equation and slopes of the three curves the required inequalities are readily proved.

If weights ($w = a + b$) proportional to the average of the positive fundamentals (a and b) are given, the unit harmonic mean (UH) of the ratios ($m = b/a$) is indicated, where the weights used are the given weights divided by $1 + m$.

$$\begin{aligned} UH &= \Sigma(mw/\overline{1+m})/\Sigma(w/\overline{1+m}) \\ &= \Sigma[(b/a)\,(a+b)/(\overline{a+b}/a)] \div \Sigma[(a+b)/(\overline{a+b}/a)] \\ &= \Sigma b/\Sigma a \end{aligned}$$

A constant factor with w will not change the result.

If weights ($w = a^{1/2}b^{1/2}$) proportional to the geometric mean of the positive fundamentals (a and b) are given, the root-harmonic mean (RH) of the ratios ($m = b/a$) is indicated, where the weights used are the given weights divided by the square root of the ratio.

$$\begin{aligned} RH &= \Sigma(mw/m^{1/2})/\Sigma(w/m^{1/2}) \\ &= \Sigma m^{1/2}w/\Sigma wm^{-1/2} \\ &= \Sigma[(b^{1/2}/a^{1/2})\ (a^{1/2}b^{1/2})] \div \Sigma[(a^{1/2}b^{1/2})/(b^{1/2}/a^{1/2})] \\ &= \Sigma b/\Sigma a \end{aligned}$$

A constant factor with w will not change the result.

The geometric mean of a geometric series, having frequencies forming a symmetrical distribution, is identical with the root harmonic mean.

This proposition may be visualized by the averaging of the accompanying m's:

m	f	wt	$m \times wt$
a	1	$1/\sqrt{a}$	$\sqrt{a}$
a^2	2	$2/\sqrt{a^2}$	$2\sqrt{a^2}$
a^3	1	$1/\sqrt{a^3}$	$\sqrt{a^3}$

$$RH = (\sqrt{a} + 2\sqrt{a^2} + \sqrt{a^3})/(1/\sqrt{a^3} + 2/\sqrt{a^2} + 1/\sqrt{a}) = a^2$$

It is obvious from the arrangement of the weights and the products that the root harmonic mean of such a series will always be the square root of the first and last m's, which is also the geometric mean.

This proposition still holds if the m series is at various ratio intervals, provided these intervals are symmetrically arranged.

Dispersion

The standard deviation (σ) of given variable items (m) about the arithmetic mean (M) is a minimum. Take a variable origin R:

Then, $\sigma^2 = \Sigma(m - R)^2 \div n = \Sigma(m^2 - 2mR + R^2) \div n$

If σ or σ^2 is a minimum,

$$d\sigma^2/dR = \Sigma 2(R - m) \div n = 0$$

Hence $\Sigma m = nR;\quad R = \Sigma m/n = M$

Prove that the square of the standard deviation is equal to the mean of the squares of the items (m) less the square of the mean (M) of the items; that is, prove $\sigma^2 = (\Sigma m^2/n) - M^2$:

By definition, expanding, and taking $M = \Sigma m/n$,

$$\begin{aligned} \sigma^2 &= \Sigma(m - M)^2 \div n = (\Sigma m^2 - 2\overline{\Sigma m}^2/n + n\overline{\Sigma m}^2/n^2) \div n \\ &= (\Sigma m^2 - \overline{\Sigma m}^2/n) \div n = (\Sigma m^2/n) - M^2 \end{aligned}$$

It is obvious that adding any constant to the items of a series will not change the standard deviation, since it changes the mean by the same constant and therefore does not change the deviations from the mean.

Index numbers

The analysis of Fisher's ideal index number may be indicated as follows:

$$Q_b \times P_r = V: \frac{\Sigma p_0 q_1}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

and similarly $Q_r \times P_b = V$

Hence, $P_i Q_i = (P_b P_r)^{1/2} (Q_b Q_r)^{1/2} = (P_b Q_r)^{1/2} (P_r Q_b)^{1/2} = V^{1/2} V^{1/2} = V$

Or the same evidence of consistency may be adduced by writing the formulas in full, thus:

$$Q_i = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_1 q_0}}$$

$$P_i = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$$

If these two formulas are multiplied, we have:

$$P_i Q_i = \sqrt{\frac{\Sigma p_1 q_1}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}} = V$$

Trends

Certain uses of the derivatives and integrals of trends have been noted. The more important derivative equations are given below. In these equations, y represents the vertical scale (trend) and x the horizontal scale (time), while u and v are functions of x. For example, u might equal $2x^2$, or $a + bx + cx^2$, which magnitudes have a defined relation to x. The letter C means a constant term, and k a constant factor. Expressions like dy/dx or du/dx mean the slopes of the curve y or u with respect to x (that is, as measured along the x horizontal scale).

Derivative equations:

$$\frac{d(x^n)}{dx} = nx^{n-1}$$

For example, the derivative (dy/dx) of the function, $y = x^3$ is $3x^2$ (slope). And of $y = x^{1/2}$ the derivative is $\frac{1}{2}x^{-1/2} = 1 \div (2\sqrt{x})$.

$$\frac{d(ky)}{dx} = k\frac{dy}{dx}$$

That is, a constant factor reappears in the derivative; for example, if $y = 5x^2$; $dy/dx = 10x$. Or, if $y = 5(x^2 + 3x)$, $dy/dx = 5(2x + 3)$.

$$\frac{dC}{dx} = 0$$

That is, if $y = C$ (a constant), the slope is zero.

$$\frac{d(u + v)}{dx} = \frac{du}{dx} + \frac{dv}{dx}$$

That is, successive terms are differentiated separately. For example,

$$\frac{d(a + bx + cx^2)}{dx} = b + 2c$$

$$\frac{d(uv)}{dx} = u\frac{dv}{dx} + v\frac{du}{dx}$$

This expresses the method of differentiating the product of two functions of x. For example, if $y = x^2(x^3 - 1)$, $dy/dx = 2x(x^3 - 1) + 3x^2 \times x^2 = 5x^4 - 2x$. This result might also be found by writing $y = x^5 - x^2$ and differentiating the terms separately.

$$\frac{d(u/v)}{dx} = \frac{(du/dx)v - u(dv/dx)}{v^2}$$

This expresses the derivative of $u \div v$, and may be used also if numerator or denominator is a constant. For example, if

$$y = \frac{x^3}{x^2 - 1},\ dy/dx = [3x^2(x^2 - 1) - x^3 \times 2x] \div (x^2 - 1)^2 = (x^4 - 3x^2) \div (x^2 - 1)^2$$

An expression like $y = 4/x^2$ may be differentiated as a fraction:

$$dy/dx = (0 \times x^2 - 4 \times 2x) \div x^4 = -\ 8/x^3$$

Or it may be written:

$$y = 4x^{-2};\quad dy/dx = -\ 8x^{-3} = -\ 8/x^3$$

The following derivative equations may be added (angles are measured in radians, 2π radians = 360°):

$$\frac{d(\sin u)}{dx} = \cos u \frac{du}{dx};\qquad \frac{d(\cos u)}{dx} = -\ \sin u \frac{du}{dx};$$

$$\frac{d(\tan u)}{dx} = \sec^2 u \frac{du}{dx};\qquad \frac{d(\cot u)}{dx} = -\ \csc^2 u \frac{du}{dx};$$

$$\frac{d(\sec u)}{dx} = \sec u \tan u \frac{du}{dx};\qquad \frac{d(\csc u)}{dx} = -\ \csc u \cot u \frac{du}{dx};$$

$$\frac{d(\log_e u)}{dx} = \frac{du/dx}{u};\qquad \frac{d(\log_{10} u)}{dx} = \frac{0.4343}{u} \times \frac{du}{dx};$$

$$\frac{dk^u}{dx} = k^u \frac{du}{dx} \log_e k;\qquad \frac{de^x}{dx} = e^x$$

$$\frac{de^u}{dx} = e^u \frac{du}{dx}$$

Integration, which expresses algebraically the area (A) under a curve, is the reverse of differentiating and may be performed by reversing the operation of the rules above. For example, if $y = 2x$; $A = x^2 + C$—the latter term is an unknown constant. If C may be disregarded; $A = x^2$ will give the area under the curve, $y = 2x$, from the origin to any specified x. The area from one specified x to another

x, may be found by substituting each x successively in the integral equation, and subtracting the areas thus found. The unknown constant, C, will not affect this result.

Prove that the derivative of the function

$$T = kx^n \text{ is } dT/dx = knx^{n-1}$$

Assuming that x is time, at x, $T = kx^n$; and a little later, at $x + \Delta x$ on the time scale,

$$\begin{aligned} T &= k(x + \Delta x)^n \\ &= kx^n + knx^{n-1}\Delta x + kn(n-1)(x^{n-2})(\Delta x)^2 \div 2 + \ldots k(\Delta x)^n \end{aligned}$$

To measure the rise (ΔT) of T between x and $x + \Delta x$, subtract

$$\Delta T = k(x + \Delta x)^n - kx^n = knx^{n-1}\Delta x + kn(n-1)(x^{n-2})(\Delta x)^2 \div 2 + \ldots {}^k(\Delta x)^n$$

Divide the equation by Δx, and diminish Δx so that ΔT and Δx simultaneously approach zero as a limit (Δ becomes the infinitesimal d), then, since terms having the factor dx approach zero,

$$dT/dx = knx^{n-1}$$

which is the slope of a tangent to the curve on the ordinate, x. Conversely, the antiderivative, or integral of knx^{n-1}, is kx^n; or of kx^n, is

$$(kx^{n+1}) \div (n + 1)$$

An undetermined constant (C) is understood with an integral, since, if the operation were again reversed, any constant, or group of constants, would be dropped, inasmuch as it does not affect the slope.

The normal equations of parabolas.—The derivation of the equations used with the straight-line and parabola trends may be briefly suggested. It will be seen that, in the general equation, $T = a + bx + cx^2$, a determines a central height; bx, a slope; and cx^2, a constant curvature. To equate these measures in the data (Y) and the trend (T), we assume that $\Sigma Y = \Sigma T$; $\Sigma xY = \Sigma xT$; and $\Sigma x^2Y = \Sigma x^2T$. If $a + bx + cx^2$ is substituted for T in each of these equalities, the so-called normal equations are obtained:

$$\begin{aligned} \Sigma Y &= na + b\Sigma x + c\Sigma x^2 \\ \Sigma xY &= a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \\ \Sigma x^2Y &= a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 \end{aligned}$$

The normal equations of the straight-line trend, similarly derived, will lack the c terms and the Σx^2Y equation, while the normal equations for the cubic will require d terms and a Σx^3Y equation, which may be added by noting the succession of the powers of x. The equations previously discussed in trend fitting are derived by simple algebraic transformation from the normal equations on the assumption that, with time centered, sums of odd powers of x equal zero. When by reason of weightings or irregular intervals this assumption does not hold true, the constants must be obtained by solving the normal equations.

That the parabola trend fitted to data by the least squares formulas has the least possible standard deviation ($\Sigma\overline{Y - T}^2$ a minimum) of any curve of its type, may be proved by the calculus method of derivation, from the equation, where $d = Y - T$,

$$\Sigma d^2 = \Sigma(Y - T)^2 = \Sigma(Y - a - bx - cx^2)^2$$

The first derivatives with respect to a, b, and c, at the minimum of Σd^2, are, $du^n/da = nu^{n-1}du/da$, where $u = \Sigma(Y - a - bx - cx^2)^2$, etc., or

$$\begin{aligned} 2\Sigma(Y - a - bx - cx^2)(-1) &= 0 \\ 2\Sigma(Y - a - bx - cx^2)(-x) &= 0 \\ 2\Sigma(Y - a - bx - cx^2)(-x^2) &= 0 \end{aligned}$$

which reduce to the three normal equations, respectively. The same proof may be applied to a straight-line or a parabola of any degree.

In analytic geometry, the parabola is generally described as a curve every point of which is equally distant from a line called the directrix, and a point not on the line called the focus. If a horizontal line is taken as the directrix and a point located one unit above this line is taken as the focus, a parabola $y = x^2/2$ is generated. This curve has positive curvature (approximately U-shaped); the anti-mode, or lowest point, is mid-way between the focus and directrix; and the X-axis and Y-axis intersect at the anti-mode. The *latus rectum* (the chord through the focus and parallel to the directrix) is two units in length; that is, it extends one unit on each side of the focus.

The modified geometric trend.—The fitting of the modified geometric curve by the method of grouped data may be explained as follows: Assume for convenience six items, with the origin at the initial item:

$$(a + b);\quad (a + bc);\quad (a + bc^2);\quad (a + bc^3);\quad (a + bc^4);\quad (a + bc^5)$$

Take the following summations (S_1, S_2, and S_3):

$$S_1 = (a + b) + (a + bc);\ S_2 = (a + bc^2) + (a + bc^3);\ S_3 = (a + bc^4) + (a + bc^5)$$

$$S_1 = 2a + b(c + 1);\ S_2 = 2a + bc^2(c + 1);\ S_3 = 2a + bc^4(c + 1)$$

Differencing these summations eliminates the a's and dividing the differences leaves only c^2 (or in general c^m); hence $c^2 = (S_3 - S_2)/(S_2 - S_1) = c^m$. And

$$\begin{aligned} S_2 - S_1 &= b(c^2 - 1)(c + 1) = b(c^2 - 1)^2/(c - 1), \text{ since } (c^m - 1)/(c - 1) \\ &= c^{m-1} + c^{m-2} \ldots c^{m-m} \end{aligned}$$

or, in general, $$(S_2 - S_1) = b(c^m - 1)^2/(c - 1)$$

hence $$b = (S_2 - S_1)(c - 1)/(c^m - 1)^2$$

Also $$(S_2 - S_1)/(c^m - 1) = b(c + 1) = S_1 - 2a$$

and $$a = [S_1 - (S_2 - S_1)/(c^m - 1)] \div m$$

When a modified geometric trend is to be passed through only three points, Y_1, Y_2, and Y_3, separated by t units,

$$Y_1 = a + b;\quad Y_2 = a + bc^t;\quad \text{and} \quad Y_3 = a + bc^{2t}$$

Differencing these items eliminates the a's, and the ratio of the differences is c^t; hence

$$c^t = (Y_3 - Y_2)/(Y_2 - Y_1)$$

and $$Y_2 - Y_1 = b(c^t - 1);\quad \text{and} \quad b = (Y_2 - Y_1)/(c^t - 1)$$

Also $$a = Y_1 - b$$

The Pearl-Reed trend.—The equation usually given for the Pearl-Reed curve is:

$$y = k/(1 + e^{a+bx})$$

the first derivative of which is

$$dy/dx = by(k - y)/k$$

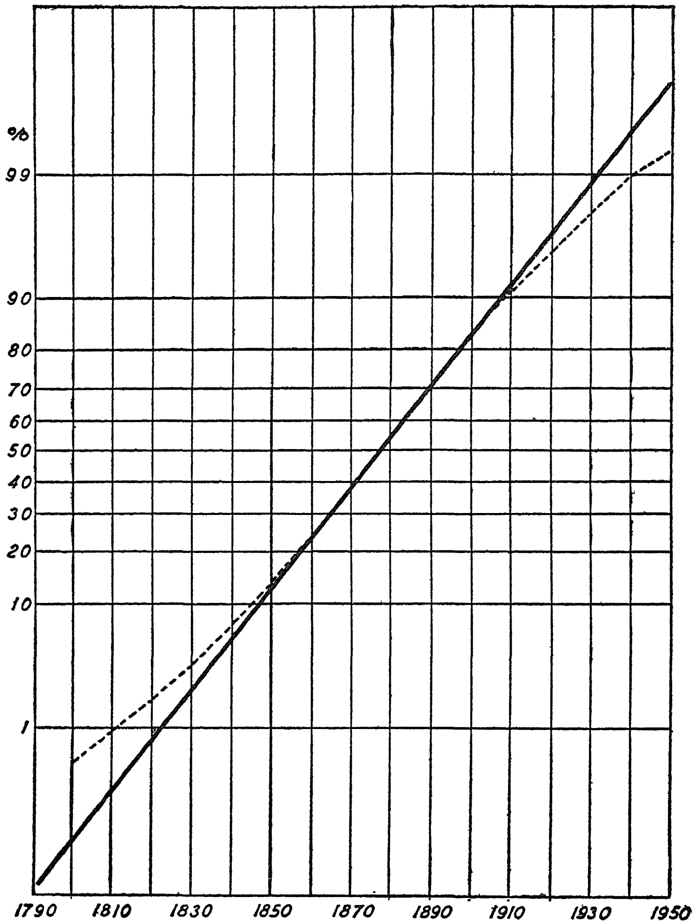


CHART 42

Pearl-Reed curve (dotted line) and normal cumulative curve (solid line) plotted on probability paper. The Pearl-Reed curve is the trend computed in Example 50. The normal curve is drawn so as to coincide with the Pearl-Reed curve in its central portion. The cumulatives of the normal cumulative curve have been read from the chart and their first differences are given in Example 89, column N. For a comparison of this normal curve with the smoothed derivative of the Pearl-Reed curve (dT/dx of Example 89), see Chart 43.

The term $e^{a+bx} = e^a \times (e^b)^x$; let

$$e^a = v \quad \text{and} \quad (e^b) = w$$

then

$$y = k/(1 + vw^x) = 1/(1/k + \overline{v/k}\, w^x)$$

The equation therefore involves three constants: $1/k$; v/k; and w, which may be written as a, b, and c, respectively. Then, $y = 1/(a + bc^x)$ or $1/y = a + bc^x$. But the original equation may prove more convenient for integration or differentiation.

The first and second derivatives of the Pearl-Reed and Gompertz curve, and the determination of each abscissa and ordinate at the respective points of inflection, may be obtained by the steps outlined as follows:

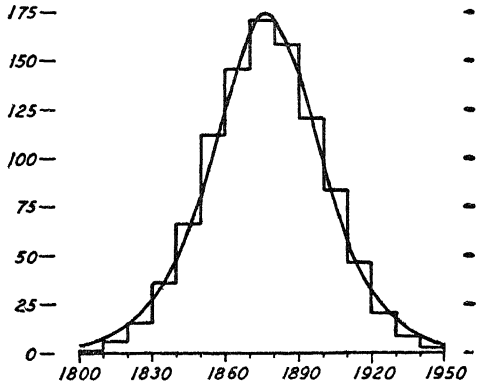


CHART 43

Comparison of a normal frequency curve of distribution, rectangular figure, and a derivative distribution based on the Pearl-Reed curve, smooth line (cf. Example 89, columns N and dT/dx). The curves were constructed graphically (cf. Chart 42) so as to coincide in their central portion. Comparison indicates that the Pearl-Reed derivative curve, thus adjusted to the normal, falls somewhat below the normal at a distance of one standard deviation from the mode, coincides with it at approximately two standard deviations, but rises above it at three standard deviations. The mode is about 1877, and one standard deviation is about 23 years.

The Pearl-Reed curve.

$$T = \frac{1}{a + bc^x}$$

$$\frac{dT}{dx} = \frac{-bc^x \log_e c}{1/T^2} = - T^2bc^x \log_e c$$

$$\frac{d(dT/dx)}{dx} = - b \log_e c \left(2T\frac{dT}{dx}c^x + T^2c^x \log_e c\right)$$

$$= b \log_e c \,(2T^3bc^{2x} \log_e c - T^2c^x \log_e c)$$

$$= c^xT^2b \log_e^2 c \,(2Tbc^x - 1)$$

Equating the second derivative to zero to find the point (x) of inflection, we have

$$c^xT^2b \log_e^2 c \,(2Tbc^x - 1) = 0$$

$$2Tbc^x = 1; \quad \text{and} \quad bc^x = \frac{1}{2T}$$

At the point of inflection,

$$T = \frac{1}{a + 1/(2T)} = \frac{1}{2a} = \frac{1}{2} \text{ of } \frac{1}{a}$$

which, when the original equation describes a Pearl-Reed curve, places the point of inflection at 50% of the upper asymptote.

The computation of a Pearl-Reed derivative distribution, and a normal curve adjusted graphically to it, is illustrated in Example 89, and Charts 42 and 43.

Example 89.—Computation of the derivative distribution based upon the Pearl-Reed curve as obtained in Example 50, and extrapolated (1800–1950), so as to approximate the completed curve. The general equation of the trend is: $T = 1 \div (a + bc^x)$, which as here modified and solved for the constants becomes: $T = 10^5 \div (100 + 1280 \times 0.5^x)$. The first derivative of the general Pearl-Reed equation is $dT/dx = -\, T^2bc^x \log_e c$, which here becomes $dT/dx = -\, (T^2bc^x \log c) \div (0.4343 \times 10^5)$, since $\log_e = \log_{10} \div 0.4343$. A frequency curve has been obtained from the trend by taking its first differences ($\Delta_1 T$). A normal frequency curve (N), adjusted to the derivative curve in its central portion, has been graphically calculated, cf. Chart 42. The derivative and first differences curves are alike, and the normal curve conforms rather closely to the two.

Year	x	Y	bc^x	T	dT/dx	$\Delta_1 T$	N
1800	−4	...	20,480	4.9	3.4		
1805	...	...	....	.....	.....	4.8	1.8
1810	−3	...	10,240	9.7	6.6		
1815	...	...	....	.....	.....	9.5	5.7
1820	−2	...	5,120	19.2	13.0		
1825	...	...	....	.....	.....	18.4	15.2
1830	−1	...	2,560	37.6	25.1		
1835	...	...	....	.....	.....	34.9	35.3
1840	0	76	1,280	72.5	46.6		
1845	...	...	....	.....	.....	62.6	67.0
1850	1	130	640	135.1	81.0		
1855	...	...	....	.....	.....	103.0	112.5
1860	2	220	320	238.1	125.7		
1865	...	...	....	.....	.....	146.6	146.6
1870	3	396	160	384.6	164.0		
1875	...	...	....	.....	.....	171.0	171.0
1880	4	530	80	555.6	171.1		
1885	...	...	....	.....	.....	158.7	158.7
1890	5	725	40	714.3	141.5		
1895	...	...	....	.....	.....	119.1	121.7
1900	6	838	20	833.3	96.3		
1905	...	...	....	.....	.....	75.7	84.0
1910	7	892	10	909.1	57.3		
1915	...	...	....	.....	.....	43.3	46.5
1920	8	959	5	952.4	31.4		
1925	...	...	....	.....	.....	23.3	20.7
1930	9	...	2.5	975.6	16.5		
1935	...	...	....	.....	.....	13.0	9.0
1940	10	...	1.25	988.6	8.5		
1945	..	...	....	.....	.....	5.2	2.7
1950	11	...	0.625	933.8	4.3		

The Gompertz curve.—A similar analysis of the Gompertz curve in its elementary form is as follows:

$$T = ab^{c^x}$$

$$\log_e T = \log_e a + c^x \log_e b$$

$$(dT/dx)/T = c^x \log_e b \times \log_e c$$

$$dT/dx = Tc^x \log_e b \times \log_e c$$

$$d(dT/dx)/dx = (Tc^x \log_e c + c^{2x}T \log_e b \log_e c) \log_e b \times \log_e c$$

$$= c^xT \log_e b \times \log_e^2 c\,(1 + c^x \log_e b)$$

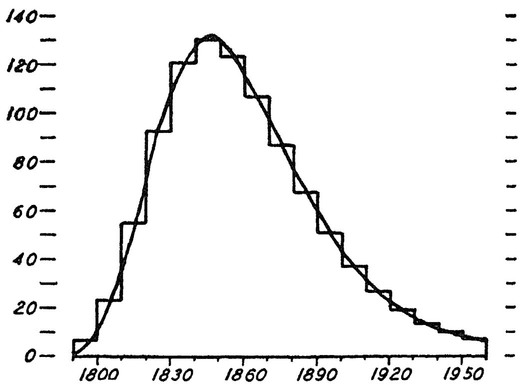


CHART 44

Comparison of a normal frequency distribution, rectangular figure, and a smoothed-line derivative distribution based on the Gompertz curve (cf. Example 90, columns N and dT/dx). Since the original Gompertz curve has its points of inflection at $a/e = 1000/2.7183$, the derivative curve is skewed in a form closely approximating a logarithmic normal. The position of the x-scale, on which the derivative curve approximates a normal, may be found by means of the quartiles as follows: Add to each x the correction: $c = (Q_2^2 - Q_1Q_3) \div (Q_1 + Q_3 - 2Q_2)$. The quartiles computed on the x-scale are $Q_1 = 3.088$; $Q_2 = 5.062$; and $Q_3 = 7.517$, and $c = (25.624 - 23.212) \div 0.481 = 5.015$, or $c = 5$, approximately.

Equating the second derivative to zero to obtain the point (x) of inflection we have

$$c^xT \log_e b \log_e^2 c\,(1 + c^x \log b) = 0$$

$$c^x \log b = -1; \quad \text{and} \quad c^x = -\frac{1}{\log_e b}$$

At point of inflection,

$$T = ab^{-\frac{1}{\log_e b}},$$

$$\log_e T = \log_e a - \frac{1}{\log_e b} \log_e b = \log_e a - 1$$

$$T = a/e$$

which when the original equation describes a Gompertz curve, places the point of inflection at 37% of the upper asymptote.

For purposes of analysis, the equation of the Gompertz curve, $T = ae^{-e^{-bx}}$, will be found convenient.

The calculation of a Gompertz derivative curve and a logarithmic normal curve adjusted graphically to it, is illustrated in Example 90 and Chart 44.

Example 90.—Computation of the derivative distribution based upon the Gompertz curve as obtained in Example 51, and extrapolated (1790–1960) so as to approximate the completed curve. The general equation of the trend is $T = ab^{c^x}$, and its first derivative is $dT/dx = Tc^x \log_e b \times \log_e c$, which here becomes $Tc^x \times \log b\ (-0.1549 \div 0.4343^2) = Tc^x \log b\ (-0.82127)$, since $\log_e = \log_{10} \div 0.4343$. A frequency curve may be obtained from the trend by taking its first differences $\Delta_1 T$. A normal frequency curve (N) adjusted to the derivative curve has been graphically calculated in a manner similar to that of Chart 42, but the x-scale was corrected by adding $c = 5$ and was written logarithmically. The normal curve thus found conforms so closely to the $\Delta_1 T$ curve that only slight divergencies at the extremes can be approximated (*cf.* Chart 44). $c = 0.7$; $\log b = -1.525$.

Year	x	Y	$c^x \log b$	T	dT/dx	$\Delta_1 T$	N
1790	−2	...	−3.1123	0.77	2.0		
1795	...	...	....	......	.....	5.9	5.5
1800	−1	...	−2.1786	6.63	11.8		
1805	...	...	....	......	.....	23.2	23.9
1810	0	29	−1.5250	29.85	37.5		
1815	...	...	....	......	.....	55.8	55.8
1820	1	90	−1.0675	85.61	75.1		
1825	...	...	....	......	.....	93.3	93.3
1830	2	174	−0.7473	178.94	109.8		
1835	...	...	....	......	.....	120.9	120.9
1840	3	302	−0.5231	299.85	128.8		
1845	...	...	....	......	.....	130.5	130.5
1850	4	450	−0.3662	430.33	130.0		
1855	...	...	....	......	.....	123.9	123.9
1860	5	532	−0.2563	554.24	116.6		
1865	...	...	....	......	.....	107.4	107.4
1870	6	641	−0.1794	661.61	97.8		
1875	...	...	....	......	.....	87.2	87.2
1880	7	770	−0.1256	748.86	77.3		
1885	...	...	....	......	.....	67.9	67.9
1890	8	820	−0.0879	816.77	59.15		
1895	...	...	....	......	.....	51.2	51.2
1900	9	848	−0.0615	867.96	43.8		
1905	...	...	....	......	.....	37.6	37.6
1910	10	921	−0.0431	905.52	32.0		
1915	...	...	....	......	.....	27.3	27.3
1920	11	935	−0.0302	932.82	23.2		
1925	...	...	....	......	.....	19.8	19.8
1930	12	...	−0.0211	952.58	16.5		
1935	...	...	....	......	.....	13.9	13.9
1940	13	...	−0.0148	966.49	11.7		
1945	...	...	....	......	.....	9.8	10.0
1950	14	...	−0.0104	976.34	8.4		
1955	...	...	....	......	.....	7.0	7.0
1960	15	...	-0.00728	983.33	5.9		

Fourier analysis.—To a limited extent the Fourier analysis has been applied to curve fitting in frequency distributions and cycles. Various methods of fitting the curve, including certain short-cuts, may be found in text books on mathematics and engineering. A simple method is as follows:

1. In the case of frequency distributions, arrange the x-scale to extend from 0 to 180, combining any very small classes at the extremes or disregarding them entirely. In the case of a cycle, draw the x-scale through the center (AM) of the series, and scale it from 0 to 360, covering the cycle which is assumed to repeat.

2. Measure at least one ordinate (Y) for every ten units on the x-scale, or more if harmonics higher than the fourth are to be fitted. If the data are incomplete, ordinates may be interpolated. Ordinarily, four harmonics will be found sufficient.

3. The equation of the Fourier series is,

$$T = A_0 + A_1 \sin x + B_1 \cos x + A_2 \sin 2x + B_2 \cos 2x, \text{ etc.}$$

In this equation A_0 is the axis of symmetry; that is, the base line in the case of the frequency distribution, or the central line in the case of the cycle. A_1 and B_1 are the coefficients of the first harmonic; A_2 and B_2 are the coefficients of the second harmonic, etc. In the case of the cycle, if it is assumed to be symmetrical, only odd numbered harmonics should be calculated, and this is often true of distributions.

4. A_1 is found by multiplying each ordinate (Y) by $\sin x$, and dividing the sum of these products by $0.5n$ (n = the number of ordinates); B_1 is similarly found by the use of $\cos x$. Harmonics of a higher degree are similarly found, but, as may be seen from the formula, the second harmonic employs the sine and cosine of $2x$, etc. Values of x running above 360 are reduced by subtracting 360 as many times as possible, leaving a positive remainder. In reading the sine and the cosine from a table, multiples of 90° are similarly disregarded except as they determine the algebraic sign. From 0° to 90°, both sine and cosine are positive; from 90° to 180° the sine is positive and the cosine negative; from 180° to 270° both sine and cosine are negative; and from 270° to 360° the sine is negative, and the cosine positive. Intermediate values may be read from a table giving sines and cosines from 0° to 90°.

In distributions the curve may be centered by taking $A_0 = (\Sigma Y - \Sigma T)/n$ or by multiplying each T by $\Sigma Y/\Sigma T$.

Time Series Analysis

In the link-relative method, logarithmic chaining of the link-relatives will produce identical seasonal indexes with any assumed base as the origin of the chaining.

Assume s subdivisions to the year, and the link-relatives whose logarithms are $r_1, r_2, r_3 \ldots r_s$. Since the logarithmic crude chain is an arithmetic cumulative of the r's, their logarithms are themselves the first differences of the logarithmic chain. But the correction for slope is Σr at the s items, counting in a circle from the assumed base, and the first differences of the corrections are the constant $\Sigma r/s$. Hence the $r - \Sigma r/s$ represents the first differences of the chain corrected for trend. Their total being zero ($\Sigma r - s\Sigma r/s$), they are repetitive, and retain the same first differences when cumulated in a circle from any base. Hence, when finally centered by adding a constant log, they are identical from any base.

To prove that the *Annalist's* method of averaging percentage cycles to obtain a composite percentage cycle (CC_p) is equivalent to the usual method of averaging

average deviation cycles, (d/AD), assuming a weighted harmonic mean of the average deviations to represent the composite average deviation:

Proof: The *Annalist's* method may be expressed as follows:

$$CC_p = \Sigma(w\,\overline{1+d}/AD)/\Sigma(w/AD) = 1 + \Sigma(dw/AD)/\Sigma(w/AD)$$

Subtract 1 to obtain a composite deviation cycle (CC_d):

$$CC_p = \Sigma(dw/AD)/\Sigma(w/AD)$$

The weighted harmonic mean (H) of the average deviations, using the same weights (w) as are used in averaging the d/AD cycles, is

$$H = \Sigma w/\Sigma(w/AD)$$

Dividing CC_d by H to obtain the d/AD composite cycle (CC),

$$\begin{aligned} CC &= [\Sigma(dw/AD)/\Sigma(w/AD)] \div [\Sigma w/\Sigma(w/AD)] \\ &= \Sigma(dw/AD)/\Sigma w \end{aligned}$$

which is identical with the weighted average of the d/AD cycles.

Correlation

The introductory formulas for the coefficient of similarity (Sm) and linear correlation (r) as illustrated on AD or σ scatter diagrams in the form of Chart 33*a*, p. 230, may be derived as follows. The diagonal deviations are

$$d_w = (y - x)/\sqrt{2}$$

$$d_v = (y + x)/\sqrt{2}$$

and

$$Sm = (AD_v - AD_w)/\sqrt{2} = \Sigma s/n$$

where s is the smaller of each two paired x/AD_x and y/AD_y correlatives, written with the sign of the correlation indicated by that pair (like signs give $+$; unlike $-$). The steps in the foregoing transformation are given in an article, "First Moment Correlation," *Journal of the American Statistical Association*, December, 1930.

Also,

$$r = (\sigma_v^2 - \sigma_w^2)/2 = \Sigma xy/n$$

(by expanding and combining). But x and y are in σ units respectively; hence in ordinary deviations, x and y,

$$r = \Sigma x_\sigma y_\sigma/n = \Sigma xy/n\sigma_x\sigma_y$$

(*Note*: σ or AD as a subscript to a variable is understood to be the σ or AD of the variable.)

The formula may be conveniently written (see below)

$$r = \Sigma xy/\sqrt{\Sigma x^2 \Sigma y^2}$$

The coefficient of correlation $r = \Sigma xy/\sqrt{\Sigma x^2\Sigma y^2}$ as applied to the paired deviations x and y, may be written so that it can be applied to the data X and Y, assuming that the deviations are to be taken from the respective means, as follows:

$$r = (\Sigma XY - nM_xM_y)/[(\Sigma X^2 - nM_x^2)\ (\Sigma Y^2 - nM_y^2)]^{1/2}$$

in which M_x and M_y are the arithmetic means of the X and Y series, respectively.

This transformation is readily made by substituting in the equation first given the following equalities:

$$x = X - M_x; \quad y = Y - M_y$$

expanding and combining the terms.

The relation of the coefficient of correlation (r) to the slope (b) of the regression (trend) line of Y on X, is as follows.

In units of centered deviations, x and y,

$$\begin{aligned} r &= \Sigma xy/n\sigma_x\sigma_y \\ &= \Sigma xy/n(\Sigma x^2/n)^{1/2}(\Sigma y^2/n)^{1/2} = \Sigma xy/(\Sigma x^2\Sigma y^2)^{1/2} \\ b &= \Sigma xy/\Sigma x^2; \quad r/b = (\Sigma x^2)^{1/2}/(\Sigma y^2)^{1/2} = \sigma_x/\sigma_y \\ r &= b\sigma_x/\sigma_y \quad \text{and} \quad b = r\sigma_y/\sigma_x \end{aligned}$$

The coefficient of linear correlation (r) is equivalent to $(1 - S^2/\sigma_y{}^2)^{1/2}$ where S is the standard error of estimate (standard deviation of the Y's from the regression line or trend, T).

Write the Y's in σ_y units, and the X's in σ_x units from their respective means.

$$T = x\Sigma xy/\Sigma x^2$$

The deviations (d) of the y's from T are

$$d = y - x\Sigma xy/\Sigma x^2$$

and the standard error of estimate (S) is given by

$$\begin{aligned} nS^2 &= \Sigma(y - x\Sigma xy/\Sigma x^2)^2 \\ &= \Sigma y^2 - 2(\Sigma xy)^2/\Sigma x^2 + (\Sigma x^2)\,(\Sigma xy)^2/(\Sigma x^2)^2 \end{aligned}$$

whence $$S^2 = [\Sigma y^2 - (\Sigma xy)^2/\Sigma x^2]/n = 1 - (\Sigma xy/\Sigma x^2)^2$$

and $$1 - S^2 = 1 - 1 + (\Sigma xy/\Sigma x^2)^2 = r^2$$

or, in ordinary x and y deviation units,

$$r^2 = 1 - S^2/\sigma_y{}^2$$

If r is known, S may be calculated by solving the preceding equation for S

$$S^2 = \sigma_y{}^2\,(1 - r^2)$$

As applied to serial rankings of correlated data X and Y,

$$r_r = 1 - [6\Sigma d^2/n(n^2 - 1)] = r$$

where d is the difference between any two paired ranks, and n is the number of items in each series.

Note that the two serial rankings, X and Y, have the same average; that therefore $X - Y = x - y$ or the ranks centered; and $\Sigma x^2 = \Sigma y^2$, and $\sigma_x = \sigma_y$, also that $\Sigma x^2 = n(n^2 - 1)/12$; (cf. *Journal of the American Statistical Association*, March, 1925); then, by successive transformations

$$\begin{aligned} r_r &= 1 - [6\Sigma d^2/n(n^2 - 1)] \\ &= 1 - \Sigma(x - y)^2/2\Sigma x^2 \\ &= 1 - (\Sigma x^2 - 2\Sigma xy + \Sigma y^2)/2\Sigma x^2 \\ &= 1 - (2\Sigma x^2 - 2\Sigma xy)/2\Sigma x^2 \\ &= \Sigma xy/\Sigma x^2 \end{aligned}$$

which is the slope (b) of the regression line or $r\sigma_y/\sigma_x$ which here equals r, since $\sigma_x = \sigma_y$.

If a trend of the correlation of Y on X consists of the average of the Y-columns, (M), as in the correlation ratio (η), then $\Sigma YT = \Sigma T^2$ and $\eta^2 = \sigma_t^2/\sigma_y^2$.

Assume n items; n_1 in the Y_1 column, n_2 in the Y_2 column, etc. Then the trend (T) is

$$T = \Sigma Y_1/n_1; \quad \Sigma Y_2/n_2, \text{ etc.}$$

the average of which (weights n_1, n_2, etc.) is $\Sigma Y/n = M$

$$\Sigma YT = \Sigma Y_1(\Sigma Y_1/n_1) + \Sigma Y_2(\Sigma Y_2/n_2), \text{ etc.}$$

$$\Sigma T^2 = n_1(\Sigma Y_1/n_1)^2 + n_2(\Sigma Y_2/n_2)^2, \text{ etc.}$$

which are obviously equal.

Hence the correlation ratio may be calculated as (cf. ρ^2 below).

$$\eta^2 = 1 - S^2/\sigma_y^2 = \sigma_t^2/\sigma_y^2$$

When a correlation table is employed, Σxy is computed from X and Y series, each centered at an assumed rather than a true average. The amount of correction required may be determined as follows: Let x and y be deviations from true averages (AM). Applying constants c_x and c_y, we have:

$$\begin{aligned} \Sigma(x + c_x)(y + c_y) &= \Sigma(xy + xc_y + yc_x + c_xc_y) \\ &= \Sigma xy + c_y\Sigma x + c_x\Sigma y + nc_xc_y \\ &= \Sigma xy + nc_xc_y \text{ (since } \Sigma x = 0 = \Sigma y) \end{aligned}$$

It follows that when Σxy is computed from uncentered x and y series, it is too large by n times the product of the corrections required to adjust the assumed averages, or nc_xc_y. The values of c_x and c_y will be found in computing σ_x and σ_y. The formula then becomes

$$\begin{aligned} r &= (\Sigma xy - nc_xc_y)/n\sigma_x\sigma_y \\ &= (\Sigma xy/n - c_xc_y)/(\sigma_x\sigma_y) \end{aligned}$$

Prove the identity, under certain conditions, of (1) the general coefficient of correlation (ρ); (2) the linear correlation of the data and a trend fitted to the data (r_{yt}); and (3) the ratio of the standard deviation of the trend (T) to the standard deviation of the data (Y) (σ on Y-scale).

Write M as arithmetic mean of Y and its trend, and S as the standard error of estimate, then (write $M = \Sigma Y/n = \Sigma T/n$, and assume $\Sigma YT = \Sigma T^2$):

$$\begin{aligned} \sigma_t^2 &= \Sigma(T - M)^2/n = (\Sigma T^2 - nM^2)/n \\ \sigma_y^2 &= \Sigma(Y - M)^2/n = (\Sigma Y^2 - nM^2)/n \\ S^2 &= \Sigma(Y - T)^2/n = (\Sigma Y^2 - \Sigma T^2)/n \end{aligned}$$

Assume least squares straight-line and parabola trends, where $\Sigma YT = \Sigma T^2$ (proof below); then, by above:

$$\rho^2 = 1 - S^2/\sigma_y^2 = (\Sigma T^2 - nM^2)/(\Sigma Y^2 - nM^2)$$

$$r^2_{yt} = [\Sigma(Y - M)(T - M)]^2/\Sigma(Y - M)^2\Sigma(T - M)^2 = (\Sigma T^2 - nM^2)/(\Sigma Y^2 - nM^2)$$

$$\rho^2 = \sigma_t^2/\sigma_y^2 = (\Sigma T^2 - nM^2)/(\Sigma Y^2 - nM^2)$$

Hence $\rho^2 = r_{yt}^2 = \sigma_t^2/\sigma_y^2$, when $\Sigma YT = \Sigma T^2$.

The identity of $\Sigma YT = \Sigma T^2$ for a quadratic parabola may be shown thus:

$$T = a + bx + cx^2; \quad \Sigma YT = a\Sigma Y + b\Sigma xY + c\Sigma x^2 Y$$

The normal equations of the parabola are:

$$\Sigma Y = na + b\Sigma x + c\Sigma x^2$$

$$\Sigma xY = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$$

$$\Sigma x^2Y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$$

Hence ΣYT = sum of:

$$a\Sigma Y = na^2 + ab\Sigma x + ac\Sigma x^2$$

$$b\Sigma xY = ab\Sigma x + b^2\Sigma x^2 + bc\Sigma x^3$$

$$c\Sigma x^2Y = ac\Sigma x^2 + bc\Sigma x^3 + c^2\Sigma x^4$$

These tabulated summations may be written by guide factors thus:

$$\begin{array}{c|ccc} & a & bx & cx^2 \\ \hline a & na^2 & + ab\Sigma x & + ac\Sigma x^2 \\ bx & ab\Sigma x & + b^2\Sigma x^2 & + bc\Sigma x^3 \\ cx^2 & ac\Sigma x^2 & + bc\Sigma x^3 & + c^2\Sigma x^4 \end{array} \left.\vphantom{\begin{array}{c}a\\a\\a\end{array}}\right\} = \Sigma YT = \Sigma T^2$$

But this set-up obviously expresses $\Sigma T^2 = \Sigma(a + bx + cx^2)^2$ and may be contracted to a straight-line trend, or expanded to cubics, quartics, etc.

Hence, for any least squares parabola,

$$\Sigma YT = \Sigma T^2$$

The normal equations of which the coefficients and y-products are given are derived as follows: Assume two independent series (u and v) and one dependent series (y) each centered ($u = U - \Sigma U/n$, etc.) as deviations from its average. Required, a and b, such that $au + bv = y'$, where $\Sigma(y - y')^2$ is a minimum. Substituting $au + bv$ for y' in $\Sigma(y - y')^2$, differentiating successively with respect to a and b, and equating to zero, we have the normal equations:

$$a\Sigma u^2 + b\Sigma uv = \Sigma uy$$

$$a\Sigma uv + b\Sigma v^2 = \Sigma vy$$

the coefficients and y-products being readily obtained by multiplying $au + bv = y$ (instead of y') successively by u and v, and summating in each case. The equations

may be solved in this form, or they may be reduced to r values by assuming the data to be expressed as σ-cycles, thus making $\sigma = 1$. Dividing through by n, we obtain:

$$a + br_{uv} = r_{uy}$$

$$ar_{uv} + b = r_{vy}$$

which, expanded to include n independent series, is the form here used.

To prove the formulas for r involving diagonal deviations:

It may be shown, by elementary geometry, that in a squared chart of x and y deviations, the diagonal deviations d_v (upper right, lower left) and d_w (upper left, lower right) are:

$$d_v = (y + x)/\sqrt{2}$$

$$d_w = (y - x)/\sqrt{2}$$

Hence, the diagonal variances are:

$$\sigma_v{}^2 = \Sigma(y + x)^2 \div (2n)$$

$$\sigma_w{}^2 = \Sigma(y - x)^2 \div (2n)$$

and their sums and differences are:

$$\sigma_v{}^2 + \sigma_w{}^2 = \Sigma x^2/n + \Sigma y^2/n = \sigma_x{}^2 + \sigma_y{}^2$$

$$\sigma_v{}^2 - \sigma_w{}^2 = 4\Sigma xy \div (2n) = 2\Sigma xy/n$$

But if the diagonal frequencies are taken as of unit class intervals, these intervals are decreased and the standard deviations increased numerically by the ratio $\sqrt{2}$, and the variances by the ratio 2. Hence, in this case, the above sums and differences are:

$$\sigma_v{}^2 + \sigma_w{}^2 = 2\sigma_x{}^2 + 2\sigma_y{}^2$$

$$\sigma_v{}^2 - \sigma_w{}^2 = 4\Sigma xy/n$$

When the cells of a double frequency distribution are set in a frame of unit class intervals, and the diagonal frequencies are also taken as unit class intervals,

$$\tfrac{1}{4}(\sigma_v{}^2 - \sigma_w{}^2) = \Sigma xy/n$$

Dividing by $\sigma_x\sigma_y$,

$$(\sigma_v{}^2 - \sigma_w{}^2) \div (4\sigma_x\sigma_y) = \Sigma xy \div (n\sigma_x\sigma_y) = r$$

But

$$\sigma_v{}^2 + \sigma_w{}^2 = 2\sigma_x{}^2 + 2\sigma_y{}^2$$

Subtracting $2\sigma_w{}^2$ from each member,

$$\sigma_v{}^2 - \sigma_w{}^2 = 2(\sigma_x{}^2 + \sigma_y{}^2 - \sigma_w{}^2)$$

Substituting (four lines above),

$$r = (\sigma_x{}^2 + \sigma_y{}^2 - \sigma_w{}^2) \div (2\sigma_x\sigma_y)$$

In negative correlations $\sigma_v{}^2$ may be substituted for $\sigma_w{}^2$ and the sign of correlation reversed.

Probability and Curve Fitting

The formulas relating to the logarithmic normal curve are discussed in "The Analysis of Frequency Distributions," by G. R. Davies, *Journal of the American Statistical Association*, December, 1929.

The mathematical aspects of binomial and other types of probability are treated in "Mathematical Statistics," by H. L. Rietz.

Prove that the normal curve of distribution expressed as $y = e^{-x^2/2}$ has a standard deviation of unity, and that the point of inflection is at $\pm 1\sigma$.

Assume the right half curve from $x = 0$ to $x = \infty$. The area (cf. Peirce, "A Short Table of Integrals," p. 63) is $\frac{1}{2}\sqrt{2\pi}$. The standard deviation is $\sigma^2 = \Sigma(x^2f) \div n$, or, with infinitesimals,

$$\sigma^2 = \frac{2}{\sqrt{2\pi}}\int_0^\infty x^2e^{-x^2/2}dx = \frac{2}{\sqrt{2\pi}} \times \tfrac{1}{2}\sqrt{2\pi} = 1$$

Hence the standard deviation is unity.

The point on the x-scale at which the curve changes from negative to positive curvature (i.e., the point of inflection) is found by equating the second derivative of the curve to zero, and solving for x, thus:

Equation of curve..........	$y = e^{-x^2/2}$
First derivative............	$y' = - e^{-x^2/2}x$
Second derivative..........	$y'' = e^{-x^2/2}x^2 - e^{-x^2/2} = 0$
Dividing by $e^{-x^2/2}$..........	$x^2 = 1$, and $x = 1$

Hence the point of inflection is at $x = 1 = \sigma$.

In the usual tables of the normal curve and its area, it is customary so to arrange the ordinates that the total area is unity; that is, the area of the right half is 0.5. As expressed above, however, the area of the half curve is $\frac{1}{2}\sqrt{2\pi}$; or $\sqrt{2\pi} = 2.506628$ for the total curve. It is obvious that the central ordinate of $y = e^{-x^2/2}$ at $x = 0$ is unity, hence to obtain unit area in the total curve it is necessary to divide the ordinates by $\sqrt{2\pi}$. This makes the central ordinate 1 divided by 2.506628 = 0.3989, and in the curve of unit area $y = (e^{-x^2/2}) \div\sqrt{2\pi}$.

Logarithms

The use of logarithms.—Every number has a corresponding logarithm, or log, which may be obtained from a table of logarithms. Conversely, if the log of a number is given, the number (antilog) may also be obtained from a table.

A logarithm consists of two parts, an integer (positive or negative) usually called the characteristic, and a fraction in decimal form usually called the mantissa. The significance of these two parts will appear later. Example:

Logarithm	Integer or characteristic	Fraction or mantissa
2.8686	2	0.8686
0.8686−3	−3	0.8686

Numbers that are powers of 10 may be written without the use of a table. Example:

Number	Log
100	2
10	1
1	0
0.1	−1
0.01	−2

This example indicates that a log is in theory the exponent which applied to 10 equates the antilog. On the basis of this theory, the various uses of logarithms in calculations may be explained.

A. To find the log of a given number.

1. Place a mark (as subscript x) immediately *after* the first significant figure of the given number and note the number of places (positive or negative) between this mark and the decimal point. This is the log integer. Examples:

Given number	Number, marked	Characteristic
7390	7_x390	3
739	7_x39	2
7.39	7_x39	0
0.0739	0.07_x39	−2
0 00739	0.007_x39	−3

2. Disregarding the position of the decimal point, look up the given number in the margins of a log table, and write the corresponding log as given in the body of the table, prefixing a decimal point. This is the log fraction, or mantissa. Example: the mantissa of 7390; 739; 0.0739; etc. is 0.8686.

3. Combine the characteristic and mantissa. Positive characteristics precede the fraction; negative characteristics follow. Examples:

Given number	Log
7390	3.8686
739	2.8686
7.39	0 8686
0.0739	0.8686−2
0.00739	0.8686−3

In order to make the negative characteristics uniform in any problem, they are often written as a combination of positive and negative characteristics. However, in statistical work it is usually more convenient to write them as first indicated, except in the case of a log that is to be divided by a certain figure. In this case the negative integer should be this figure, or a multiple of it, and a positive integer should be prefixed to balance any change that may thus be made. Examples:

Log	Divisor	Log rewritten	Log divided
0.8686−1	2	1.8686−2	0.9343−1
0.8686−4	3	2.8686−6	0.9562−2

In some cases, as when the divisor consists of several figures, or when other complex calculations are to be made, the log with a negative integer should be reduced by subtraction to a simple negative log. Thus $0.8686 - 1 = -0.1314$; $0.8686 - 2 =$

− 1.1314; etc. In this case the final result should be changed back by subtraction to the usual form.

B. To find the antilog of a given log.

1. Disregarding the characteristic of the log, look up the mantissa in the body of a log table, and from the margins note the number corresponding to it. This is the antilog, irrespective of the position of the decimal point. Thus, given the logarithm 0.8686, the antilog figures are found to be 739, the position of the decimal point being undetermined.

2. Place a mark (as subscript x) after the first significant figure of the antilog figures thus found. Point off decimally to the right (positive) or left (negative) as many places as are indicated by the characteristic, prefixing or annexing as many ciphers as may be necessary. Example:

Log	Antilog figures	Antilog
3.8686	7_x39	7390
0.8686	7_x39	7.39
0.8686−2	7_x39	0.0739

Note: In finding the log or antilog by the foregoing method, the mark (as subscript x) following the first significant figure of the antilog may of course be omitted, provided that the position which it is used to mark is mentally noted. The mark merely indicates the position of the decimal point as it is understood to be placed in the margins of the tables.

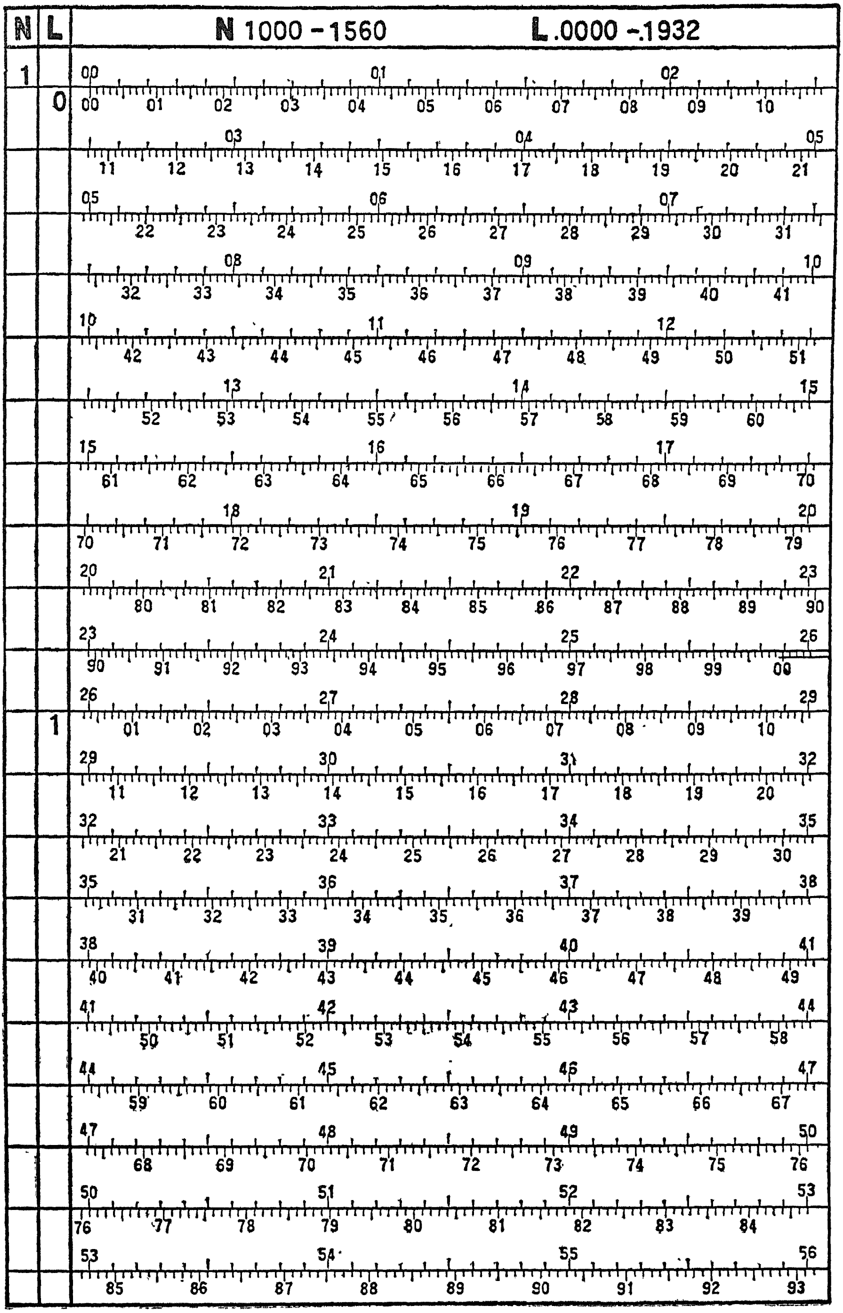
TABLE OF LOGARITHMS

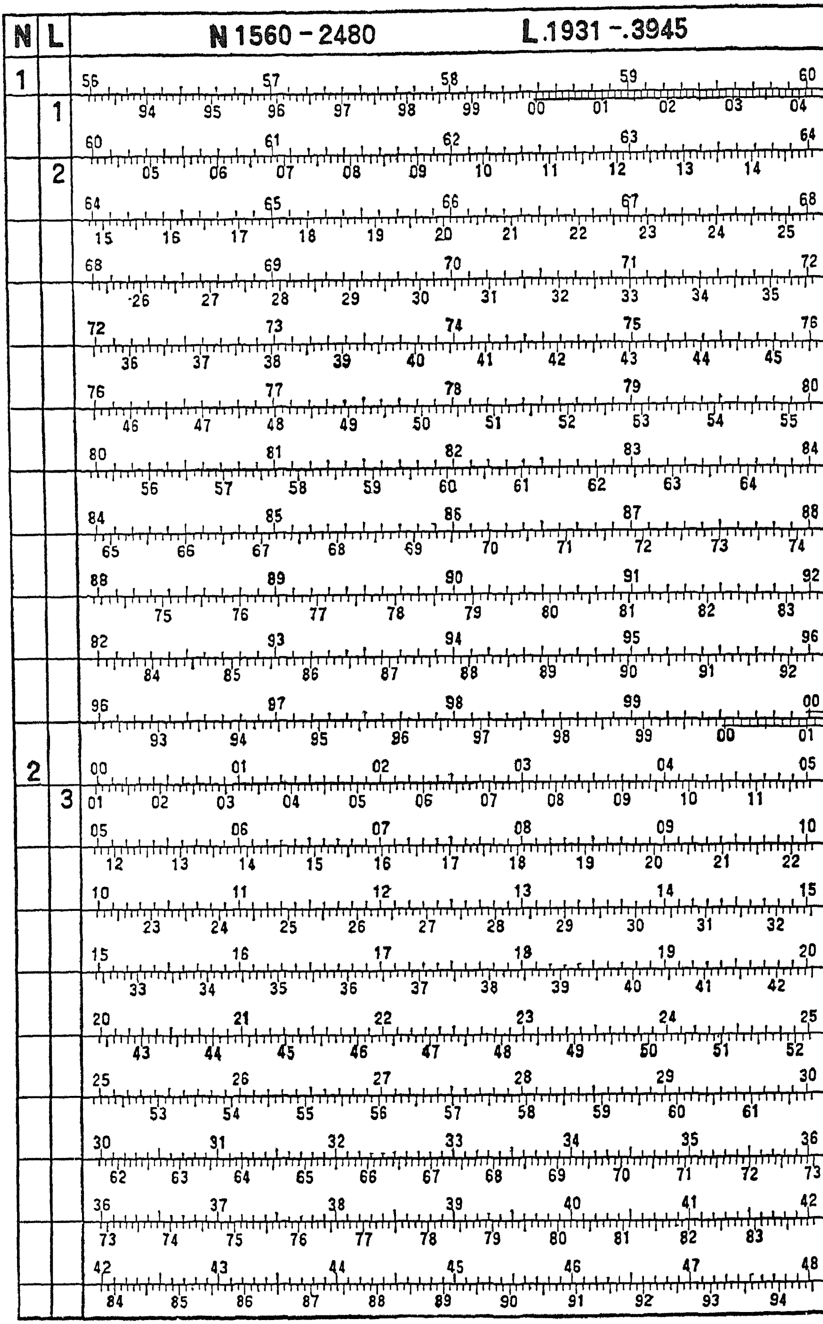
No.	0	1	2	3	4	5	6	7	8	9
1.0	0.0000	0.0043	0.0086	0.0128	0.0170	0.0212	0.0253	0.0294	0 0334	0.0374
1.1	.0414	.0453	.0492	.0531	.0569	.0607	.0645	.0682	.0719	.0755
1.2	.0792	.0828	.0864	.0899	.0934	.0969	.1004	.1038	.1072	.1106
1.3	.1139	.1173	.1206	.1239	.1271	.1303	.1335	.1367	.1399	.1430
1.4	.1461	.1492	.1523	.1553	.1584	.1614	.1644	.1673	.1703	.1732
1.5	.1761	.1790	.1818	.1847	.1875	.1903	.1931	.1959	.1987	.2014
1.6	.2041	.2068	.2095	.2122	.2148	.2175	.2201	.2227	.2253	.2279
1.7	.2304	.2330	.2355	.2380	.2405	.2430	.2455	.2480	.2504	.2529
1.8	.2553	.2577	.2601	.2625	.2648	.2672	.2695	.2718	.2742	.2765
1.9	.2788	.2810	.2833	.2856	.2878	.2900	.2923	.2945	.2967	.2989
2.0	.3010	.3032	.3054	.3075	.3096	.3118	.3139	.3160	.3181	.3201
2.1	.3222	.3243	.3263	.3284	.3304	.3324	.3345	.3365	.3385	.3404
2.2	.3424	.3444	.3464	.3483	.3502	.3522	.3541	.3560	.3579	.3598
2.3	.3617	.3636	.3655	.3674	.3692	.3711	.3729	.3747	.3766	.3784
2.4	.3802	.3820	.3838	.3856	.3874	.3892	.3909	.3927	.3945	.3962
2 5	.3979	.3997	.4014	.4031	.4048	.4065	.4082	.4099	.4116	.4133
2.6	.4150	.4166	.4183	.4200	.4216	.4232	.4249	.4265	.4281	.4298
2.7	.4314	.4330	.4346	.4362	.4378	.4393	.4409	.4425	.4440	.4456
2.8	.4472	.4487	.4502	.4518	.4533	.4548	.4564	.4579	.4594	.4609
2.9	.4624	.4639	.4654	.4669	.4683	.4698	.4713	.4728	.4742	.4757
3 0	.4771	.4786	.4800	.4814	.4829	.4843	.4857	.4871	.4886	.4900
3.1	.4914	.4928	.4942	.4955	.4969	.4983	.4997	.5011	.5024	.5038
3.2	.5051	.5065	.5079	.5092	.5105	.5119	.5132	.5145	.5159	.5172
3 3	.5185	.5198	.5211	.5224	.5237	.5250	.5263	.5276	.5289	.5302
3.4	.5315	.5328	.5340	.5353	.5366	.5378	.5391	.5403	.5416	.5428
3.5	.5441	.5453	.5465	.5478	.5490	.5502	.5514	.5527	.5539	.5551
3.6	.5563	.5575	.5587	.5599	.5611	.5623	.5635	.5647	.5658	.5670
3.7	.5682	.5694	.5705	.5717	.5729	.5740	.5752	.5763	.5775	.5786
3.8	.5798	.5809	.5821	.5832	.5843	.5855	.5866	.5877	.5888	.5899
3.9	.5911	.5922	.5933	.5944	.5955	.5966	.5977	.5988	.5999	.6010
4.0	.6021	.6031	.6042	.6053	.6064	.6075	.6085	.6096	.6107	.6117
4.1	.6128	.6138	.6149	.6160	.6170	.6180	.6191	.6201	.6212	.6222
4 2	.6232	.6243	.6253	.6263	.6274	.6284	.6294	.6304	.6314	.6325
4.3	.6335	.6345	.6355	.6365	.6375	.6385	.6395	.6405	.6415	.6425
4.4	.6435	.6444	.6454	.6464	.6474	.6484	.6493	.6503	.6513	.6522
4 5	.6532	.6542	.6551	.6561	.6571	.6580	.6590	.6599	.6609	.6618
4.6	.6628	.6637	.6646	.6656	.6665	.6675	.6684	.6693	.6702	.6712
4.7	.6721	.6730	.6739	.6749	.6758	.6767	.6776	.6785	.6794	.6803
4.8	.6812	.6821	.6830	.6839	.6848	.6857	.6866	.6875	.6884	.6893
4.9	.6902	.6911	.6920	.6928	.6937	.6946	.6955	.6964	.6972	.6981
5.0	.6990	.6998	.7007	.7016	.7024	.7033	.7042	.7050	.7059	.7067
5.1	.7076	.7084	.7093	.7101	.7110	.7118	.7126	.7135	.7143	.7152
5.2	.7160	.7168	.7177	.7185	.7193	.7202	.7210	.7218	.7226	.7235
5.3	.7243	.7251	.7259	.7267	.7275	.7284	.7292	.7300	.7308	.7316
5.4	.7324	.7332	.7340	.7348	.7356	.7364	.7372	.7380	.7388	.7396

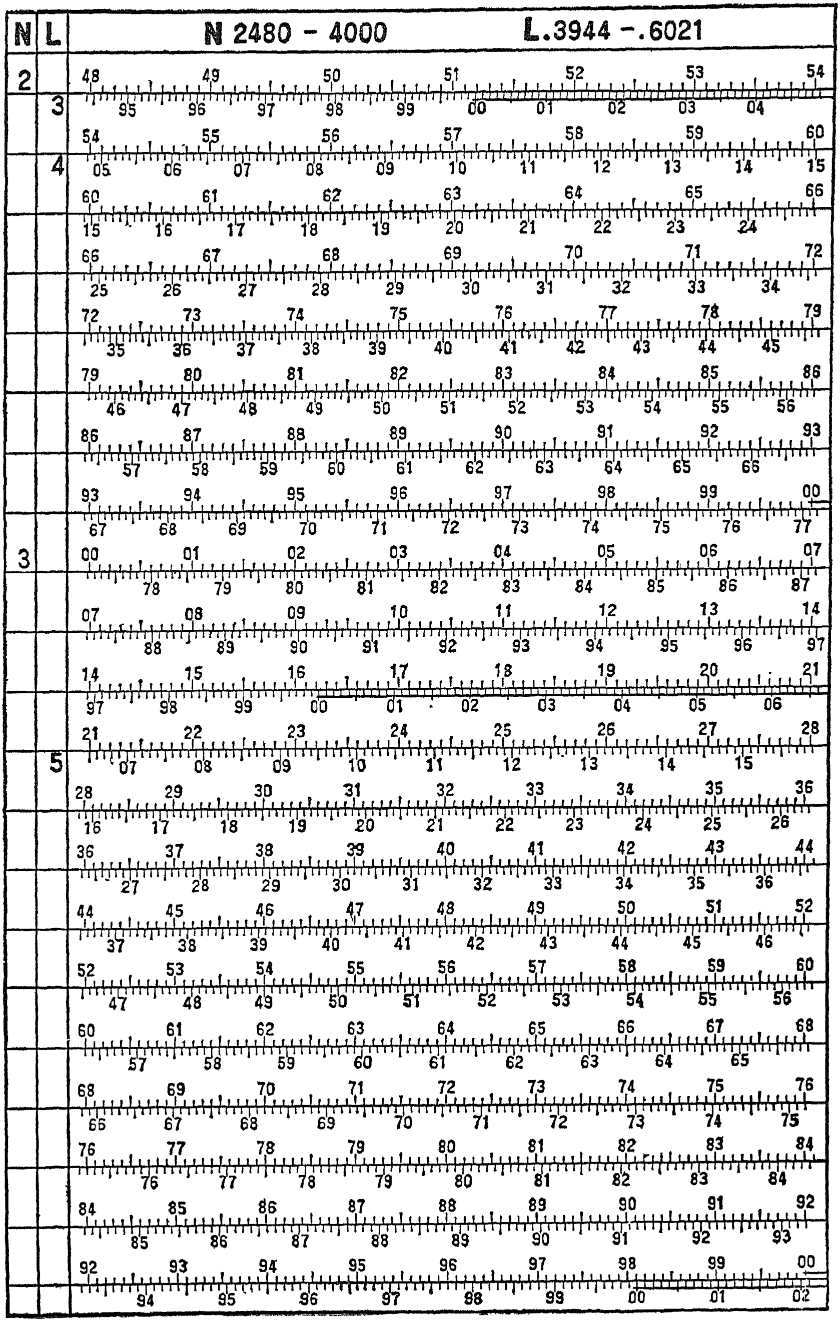
TABLE OF LOGARITHMS—*Continued*

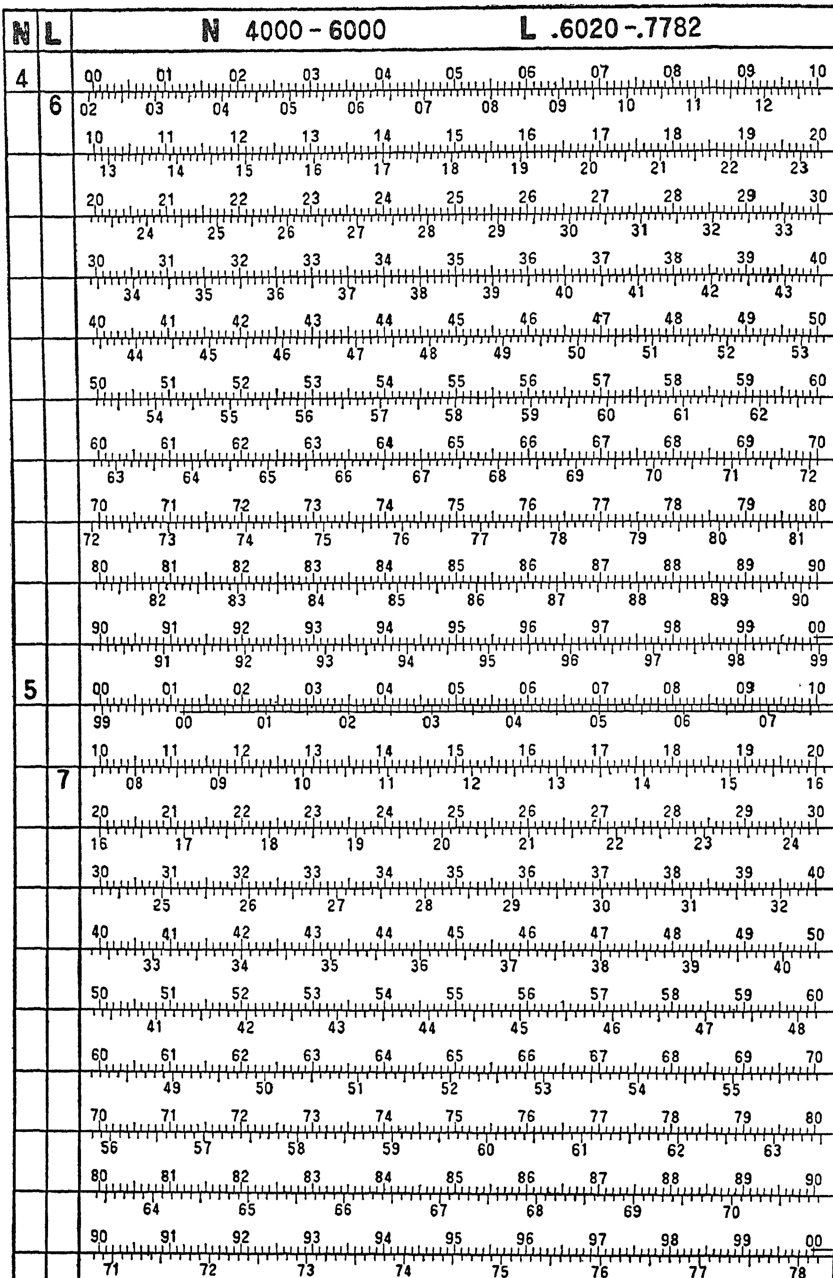
No.	0	1	2	3	4	5	6	7	8	9
5.5	0.7404	0.7412	0.7419	0.7427	0.7435	0.7443	0.7451	0.7459	0.7466	0.7474
5.6	.7482	.7490	.7497	.7505	7513	.7520	.7528	.7536	.7543	.7551
5.7	.7559	.7566	.7574	.7582	.7589	.7597	.7604	.7612	.7619	.7627
5.8	.7634	.7642	.7649	.7657	.7664	.7672	7679	.7686	.7694	.7701
5.9	.7709	.7716	.7723	.7731	.7738	.7745	.7752	.7760	.7767	.7774
6.0	.7782	.7789	.7796	.7803	.7810	.7818	.7825	.7832	.7839	.7846
6.1	.7853	.7860	.7868	.7875	.7882	.7889	.7896	.7903	.7910	.7917
6.2	.7924	.7931	.7938	.7945	.7952	.7959	.7966	.7973	.7980	.7987
6.3	.7993	.8000	.8007	.8014	.8021	.8028	.8035	.8041	.8048	.8055
6.4	.8062	.8069	.8075	.8082	.8089	.8096	.8102	.8109	.8116	8122
6.5	.8129	.8136	.8142	.8149	.8156	.8162	.8169	.8176	.8182	.8189
6.6	.8195	.8202	.8209	.8215	.8222	.8228	.8235	.8241	.8248	.8254
6.7	.8261	.8267	.8274	.8280	.8287	.8293	.8299	.8306	.8312	.8319
6.8	.8325	.8331	.8338	.8344	.8351	.8357	.8363	.8370	.8376	.8382
6.9	.8388	.8395	.8401	.8407	.8414	.8420	.8426	.8432	.8439	.8445
7.0	.8451	.8457	.8463	.8470	.8476	.8482	.8488	.8494	.8500	.8506
7.1	.8513	.8519	.8525	.8531	.8537	.8543	.8549	.8555	.8561	.8567
7.2	.8573	.8579	.8585	.8591	.8597	.8603	.8609	.8615	.8621	.8627
7.3	.8633	.8639	.8645	.8651	.8657	.8663	.8669	.8675	.8681	.8686
7.4	.8692	.8698	.8704	.8710	.8716	.8722	.8727	.8733	.8739	.8745
7.5	.8751	.8756	.8762	.8768	.8774	.8779	.8785	.8791	.8797	.8802
7.6	.8808	.8814	.8820	.8825	.8831	.8837	.8842	.8848	.8854	.8859
7.7	.8865	.8871	.8876	.8882	.8887	.8893	.8899	.8904	.8910	.8915
7.8	.8921	.8927	.8932	.8938	.8943	.8949	.8954	.8960	.8965	.8971
7.9	.8976	.8982	.8987	.8993	.8998	.9004	.9009	.9015	9020	.9025
8.0	.9031	.9036	.9042	.9047	.9053	.9058	.9063	.9069	9074	.9079
8.1	.9085	.9090	.9096	.9101	.9106	.9112	.9117	.9122	.9128	.9133
8.2	.9138	.9143	.9149	.9154	.9159	.9165	.9170	.9175	.9180	.9186
8.3	.9191	.9196	.9201	.9206	.9212	.9217	.9222	.9227	.9232	9238
8.4	.9243	.9248	.9253	.9258	.9263	.9269	.9274	.9279	.9284	.9289
8.5	.9294	.9299	.9304	.9309	.9315	.9320	.9325	.9330	.9335	.9340
8 6	.9345	.9350	.9355	.9360	.9365	.9370	.9375	.9380	.9385	9390
8.7	.9395	.9400	.9405	.9410	.9415	.9420	.9425	.9430	.9435	.9440
8.8	.9445	.9450	.9455	.9460	.9465	.9469	.9474	.9479	9484	.9489
8.9	.9494	.9499	.9504	.9509	.9513	.9518	.9523	.9528	.9533	.9538
9.0	.9542	.9547	.9552	.9557	.9562	.9566	.9571	.9576	.9581	.9586
9.1	.9590	.9595	.9600	.9605	.9609	.9614	.9619	.9624	.9628	.9633
9.2	.9638	.9643	.9647	9652	.9657	.9661	.9666	.9671	9675	.9680
9.3	.9685	.9689	.9694	.9699	.9703	.9708	.9713	.9717	.9722	9727
9.4	.9731	.9736	.9741	.9745	.9750	.9754	.9759	.9763	.9768	.9773
9.5	.9777	.9782	.9786	.9791	.9795	.9800	.9805	.9809	.9814	.9818
9.6	.9823	.9827	.9832	.9836	.9841	.9845	.9850	.9854	.9859	.9863
9.7	.9868	.9872	.9877	.9881	.9886	.9890	.9894	.9899	.9903	.9908
9.8	.9912	.9917	.9921	.9926	.9930	.9934	.9939	9943	9948	.9952
9.9	.9956	.9961	.9965	.9969	.9974	9978	.9983	.9987	.9991	.9996

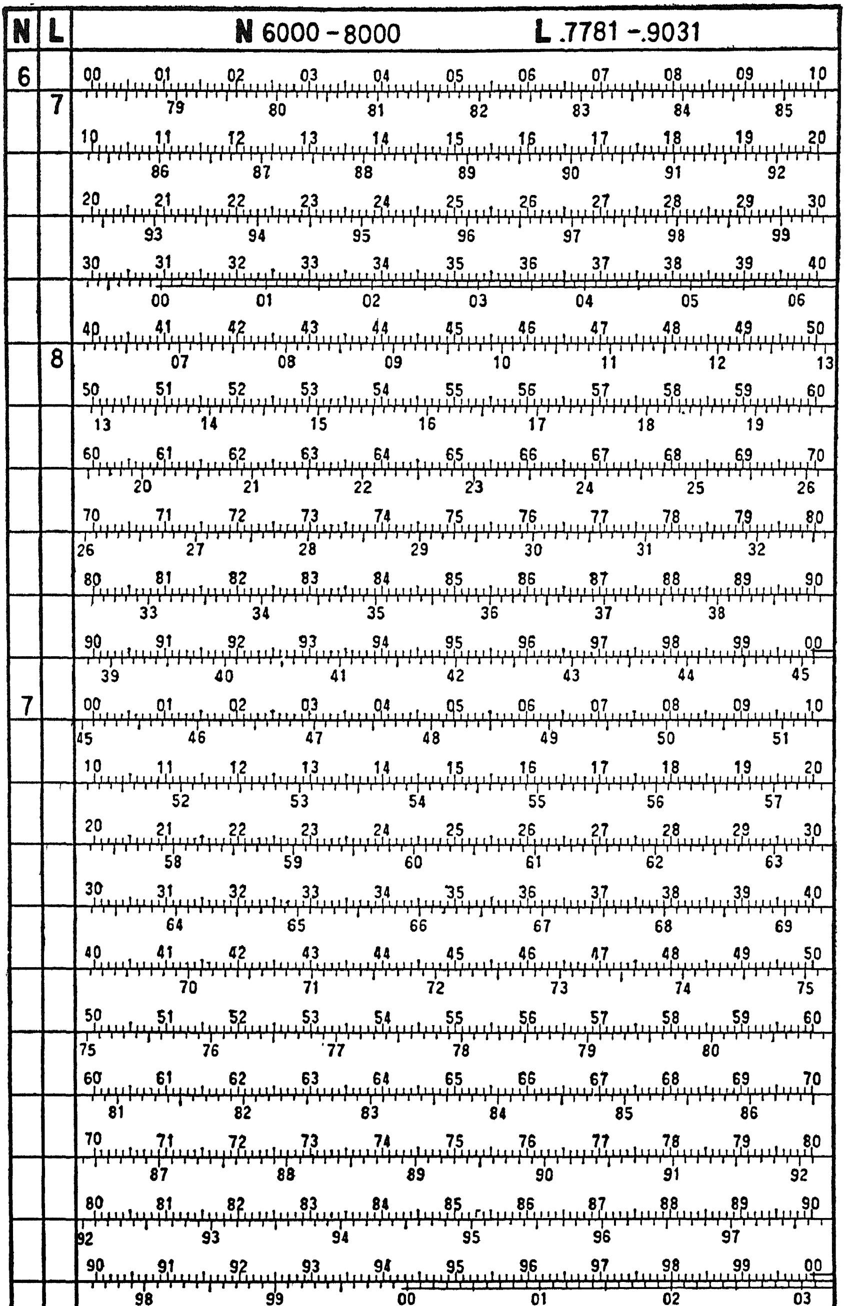
A graphic table of logarithms.—The following four-place graphic table of logarithms and antilogarithms is reprinted from Lacroix and Ragot, "A Graphic Table Combining Logarithms and Anti-Logarithms," by permission of the publishers, The Macmillan Company, New York. The first digit of the number of which the logarithm is to be taken is read in the column headed N, and succeeding figures are read in the numbers and subdivisions on the upper edge of the scale until the required point is located. The required logarithm is similarly read from the column headed L to the numbers and subdivisions below the scale, at the required point. Antilogarithms may be found by reversing this process. The rules regarding decimals and characteristics apply as before. With care, results may be read to five places. The student would do well to obtain the full-five-place table in the reference cited, as it is perhaps the most convenient and accurate table available.


N
L
N 1000 - 1560
L .0000 -.1932
1
0
1


N
L
N 1560 - 2480
L .1931 -.3945


N
L
N 2480 - 4000
L.3944 -.6021
2
3
4
5
48 49 50 51 52 53 54
95 96 97 98 99 00 01 02 03 04
54 55 56 57 58 59 60
05 06 07 08 09 10 11 12 13 14 15
60 61 62 63 64 65 66
15 16 17 18 19 20 21 22 23 24
66 67 68 69 70 71 72
25 26 27 28 29 30 31 32 33 34
72 73 74 75 76 77 78 79
35 36 37 38 39 40 41 42 43 44 45
79 80 81 82 83 84 85 86
46 47 48 49 50 51 52 53 54 55 56
86 87 88 89 90 91 92 93
57 58 59 60 61 62 63 64 65 66
93 94 95 96 97 98 99 00
67 68 69 70 71 72 73 74 75 76 77
00 01 02 03 04 05 06 07
78 79 80 81 82 83 84 85 86 87
07 08 09 10 11 12 13 14
88 89 90 91 92 93 94 95 96 97
14 15 16 17 18 19 20 21
97 98 99 00 01 02 03 04 05 06
21 22 23 24 25 26 27 28
07 08 09 10 11 12 13 14 15
28 29 30 31 32 33 34 35 36
16 17 18 19 20 21 22 23 24 25 26
36 37 38 39 40 41 42 43 44
27 28 29 30 31 32 33 34 35 36
44 45 46 47 48 49 50 51 52
37 38 39 40 41 42 43 44 45 46
52 53 54 55 56 57 58 59 60
47 48 49 50 51 52 53 54 55 56
60 61 62 63 64 65 66 67 68
57 58 59 60 61 62 63 64 65
68 69 70 71 72 73 74 75 76
66 67 68 69 70 71 72 73 74 75
76 77 78 79 80 81 82 83 84
76 77 78 79 80 81 82 83 84
84 85 86 87 88 89 90 91 92
85 86 87 88 89 90 91 92 93
92 93 94 95 96 97 98 99 00
94 95 96 97 98 99 00 01 02


N L
N 4000 - 6000
L .6020 -.7782
4
5
6
7


N L
N 6000 - 8000
L .7781 - .9031
6 | 00 01 02 03 04 05 06 07 08 09 10
7 | 79 80 81 82 83 84 85
10 11 12 13 14 15 16 17 18 19 20
86 87 88 89 90 91 92
20 21 22 23 24 25 26 27 28 29 30
93 94 95 96 97 98 99
30 31 32 33 34 35 36 37 38 39 40
00 01 02 03 04 05 06
40 41 42 43 44 45 46 47 48 49 50
8 | 07 08 09 10 11 12 13
50 51 52 53 54 55 56 57 58 59 60
13 14 15 16 17 18 19
60 61 62 63 64 65 66 67 68 69 70
20 21 22 23 24 25 26
70 71 72 73 74 75 76 77 78 79 80
26 27 28 29 30 31 32
80 81 82 83 84 85 86 87 88 89 90
33 34 35 36 37 38
90 91 92 93 94 95 96 97 98 99 00
39 40 41 42 43 44 45
7 | 00 01 02 03 04 05 06 07 08 09 10
45 46 47 48 49 50 51
10 11 12 13 14 15 16 17 18 19 20
52 53 54 55 56 57
20 21 22 23 24 25 26 27 28 29 30
58 59 60 61 62 63
30 31 32 33 34 35 36 37 38 39 40
64 65 66 67 68 69
40 41 42 43 44 45 46 47 48 49 50
70 71 72 73 74 75
50 51 52 53 54 55 56 57 58 59 60
75 76 77 78 79 80
60 61 62 63 64 65 66 67 68 69 70
81 82 83 84 85 86
70 71 72 73 74 75 76 77 78 79 80
87 88 89 90 91 92
80 81 82 83 84 85 86 87 88 89 90
92 93 94 95 96 97
90 91 92 93 94 95 96 97 98 99 00
98 99 00 01 02 03

W-5

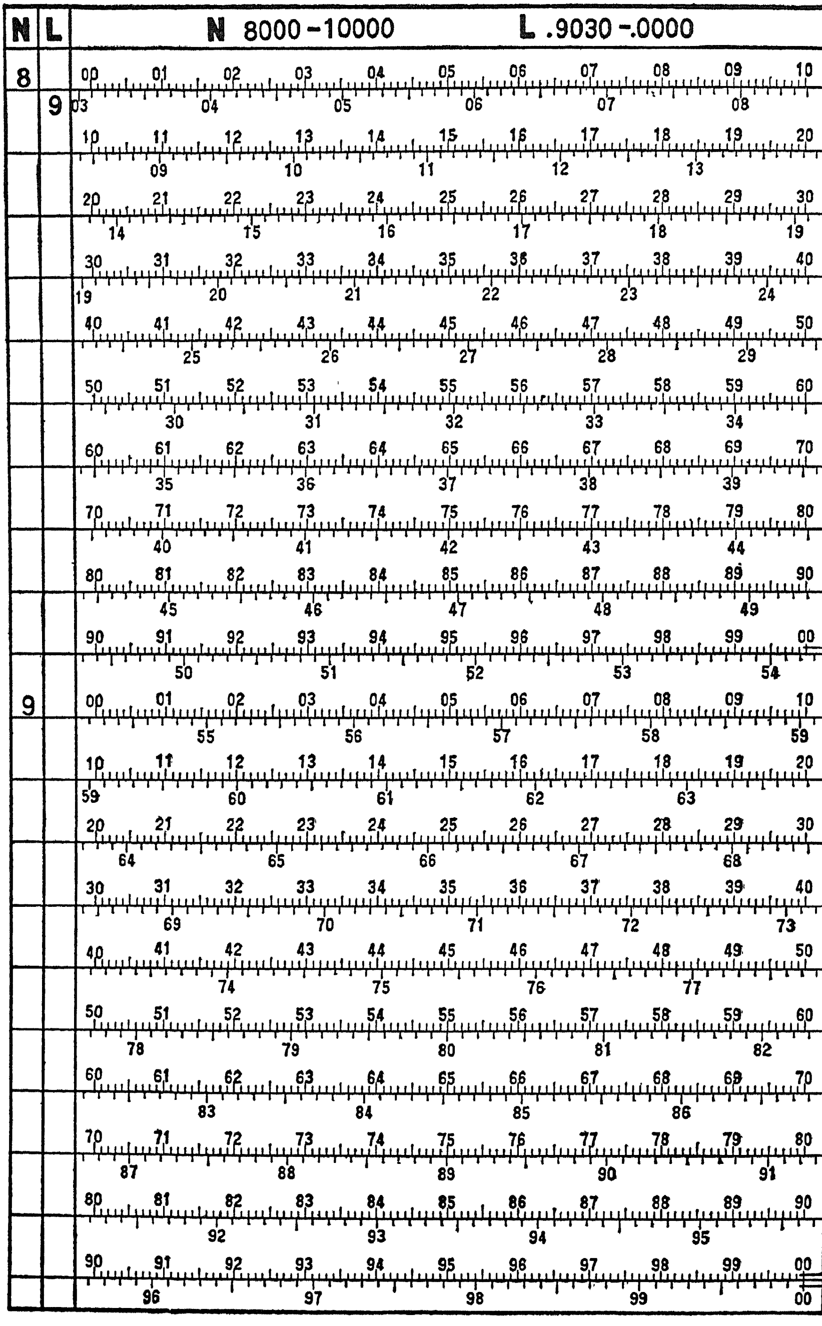

N
L
N 8000-10000
L .9030-.0000
8
9
9

TABLE OF POWERS, ROOTS, AND RECIPROCALS

No.	Square	Cube	Square root	Cube root	Reciprocals
1	1	1	1.0000	1.0000	1.0000
2	4	8	1.4142	1.2599	0.5000
3	9	27	1.7321	1.4422	0.3333
4	16	64	2.0000	1.5874	0.2500
5	25	125	2.2361	1.7100	0.2000
6	36	216	2.4495	1.8171	0.1667
7	49	343	2.6458	1.9129	0.1429
8	64	512	2.8284	2.0000	0.1250
9	81	729	3.0000	2.0801	0.1111
10	100	1,000	3.1623	2.1544	0.1000
11	121	1,331	3.3166	2.2240	0.0909
12	144	1,728	3.4641	2.2894	0.0833
13	169	2,197	3.6056	2.3513	0.0769
14	196	2,744	3.7417	2.4101	0.0714
15	225	3,375	3.8730	2.4662	0.0667
16	256	4,096	4.0000	2.5198	0.0625
17	289	4,913	4.1231	2.5713	0.0588
18	324	5,832	4.2426	2.6207	0 0556
19	361	6,859	4.3589	2.6684	0.0526
20	400	8,000	4.4721	2.7144	0.0500
21	441	9,261	4.5826	2.7589	0.0476
22	484	10,648	4.6904	2 8020	0.0455
23	529	12,167	4.7958	2.8439	0.0435
24	576	13,824	4.8990	2.8845	0.0417
25	625	15,625	5.0000	2.9240	0.0400
26	676	17,576	5.0990	2.9625	0.0385
27	729	19,683	5.1962	3.0000	0.0370
28	784	21,952	5.2915	3.0366	0.0357
29	841	24,389	5.3852	3.0723	0.0345
30	900	27,000	5.4772	3.1072	0.0333
31	961	29,791	5.5678	3.1414	0.0323
32	1,024	32,768	5.6569	3.1748	0.0313
33	1,089	35,937	5.7446	3.2075	0.0303
34	1,156	39,304	5.8310	3.2396	0.0294
35	1,225	42,875	5.9161	3.2711	0.0286
36	1,296	46,656	6.0000	3.3019	0.0278
37	1,369	50,653	6.0828	3.3322	0.0270
38	1,444	54,872	6.1644	3.3620	0.0263
39	1,521	59,319	6.2450	3.3912	0.0256
40	1,600	64,000	6.3246	3.4200	0.0250
41	1,681	68,921	6.4031	3.4482	0.0244
42	1,764	74,088	6.4807	3.4760	0.0238
43	1,849	79,507	6.5574	3.5034	0.0233
44	1,936	85,184	6.6332	3.5303	0.0227
45	2,025	91,125	6.7082	3.5569	0.0222
46	2,116	97,336	6.7823	3.5830	0.0217
47	2,209	103,823	6.8557	3.6088	0.0213
48	2,304	110,592	6.9282	3.6342	0.0208
49	2,401	117,649	7.0000	3.6593	0.0204
50	2,500	125,000	7.0711	3.6840	0.0200

Table of Powers, Roots, and Reciprocals—*Continued*

No.	Square	Cube	Square root	Cube root	Reciprocals
51	2,601	132,651	7.1414	3.7084	0.0196
52	2,704	140,608	7.2111	3.7325	0 0192
53	2,809	148,877	7.2801	3.7563	0.0189
54	2,916	157,464	7.3485	3.7798	0.0185
55	3,025	166,375	7.4162	3.8030	0.0182
56	3,136	175,616	7.4833	3.8259	0.0179
57	3,249	185,193	7.5498	3.8485	0.0175
58	3,364	195,112	7.6158	3.8709	0.0172
59	3,481	205,379	7.6811	3.8930	0.0169
60	3,600	216,000	7.7460	3.9149	0.0167
61	3,721	226,981	7.8102	3.9365	0.0164
62	3,844	238,328	7.8740	3.9579	0.0161
63	3,969	250,047	7.9373	3.9791	0.0159
64	4,096	262,144	8.0000	4.0000	0.0156
65	4,225	274,625	8.0623	4.0207	0.0154
66	4,356	287,496	8.1240	4 0412	0.0152
67	4,489	300,763	8.1854	4.0615	0.0149
68	4,624	314,432	8.2462	4.0817	0.0147
69	4,761	328,509	8.3066	4.1016	0.0145
70	4,900	343,000	8.3666	4.1213	0.0143
71	5,041	357,911	8.4261	4.1408	0.0141
72	5,184	373,248	8.4853	4.1602	0.0139
73	5,329	389,017	8.5440	4.1793	0.0137
74	5,476	405,224	8.6023	4.1983	0.0135
75	5,625	421,875	8.6603	4.2172	0.0133
76	5,776	438,976	8.7178	4.2358	0.0132
77	5,929	456,533	8.7750	4.2543	0.0130
78	6,084	474,552	8.8318	4.2727	0.0128
79	6,241	493,039	8.8882	4.2908	0.0127
80	6,400	512,000	8.9443	4.3089	0.0125
81	6,561	531,441	9.0000	4.3267	0.0123
82	6,724	551,368	9.0554	4.3445	0.0122
83	6,889	571,787	9.1104	4.3621	0.0120
84	7,056	592,704	9.1652	4.3795	0.0119
85	7,225	614,125	9.2195	4.3968	0.0118
86	7,396	636,056	9.2736	4.4140	0.0116
87	7,569	658,503	9.3274	4.4310	0.0115
88	7,744	681,472	9.3808	4.4480	0.0114
89	7,921	704,969	9.4340	4.4647	0.0112
90	8,100	729,000	9.4868	4.4814	0.0111
91	8,281	753,571	9.5394	4.4979	0.0110
92	8,464	778,688	9.5917	4.5144	0.0109
93	8,649	804,357	9.6437	4.5307	0.0108
94	8,836	830,584	9.6954	4.5468	0.0106
95	9,025	857,375	9.7468	4.5629	0.0105
96	9,216	884,736	9.7980	4.5789	0.0104
97	9,409	912,673	9.8489	4.5947	0.0103
98	9,604	941,192	9.8995	4.6104	0.0102
99	9,801	970,299	9.9499	4.6261	0.0101
100	10,000	1,000,000	10.0000	4.6416	0.0100

TABLE OF THE ORDINATES (Y) AND INTEGRAL (A), OR AREA UNDER THE CURVE OF THE RIGHT HALF OF THE NORMAL CURVE OF DISTRIBUTION,

$$Y = \frac{1}{\sqrt{2\pi}} e^{-0.5x^2},$$

where x is in units of the standard deviation, measured from the center of the curve, or mean, at $x = 0$.

x	Y	A	x	Y	A	x	Y	A
0 00	0.3989	0.0000	1.35	0.1604	0.4115	2.70	0.0104	0.4965
0.05	.3984	.0199	1.40	.1497	.4192	2.75	.0091	.4970
0.10	.3970	.0398	1.45	.1394	.4265	2.80	.0079	.4974
0.15	.3945	.0596	1.50	.1295	.4332	2.85	.0069	.4978
0.20	.3910	.0793	1.55	.1200	.4394	2.90	.0060	.4981
0 25	.3867	.0987	1.60	.1109	.4452	2.95	.0051	.4984
0 30	.3814	.1179	1.65	.1023	.4505	3.00	.0044	.4987
0.35	.3752	.1368	1.70	.0940	.4554	3.05	.0038	.4989
0.40	.3683	.1554	1.75	.0863	.4599	3.10	.0033	.4990
0.45	.3605	.1736	1.80	.0790	.4641	3.15	.0028	.4992
0.50	.3521	.1915	1.85	.0721	.4678	3.20	.0024	.4993
0.55	.3429	.2088	1.90	.0656	.4713	3.25	.0020	.4994
0.60	.3332	.2257	1.95	.0596	.4744	3.30	.0017	.4995
0.65	.3230	.2422	2.00	.0540	.4772	3.35	.0015	.4996
0.70	.3123	.2580	2.05	.0488	.4798	3.40	.0012	.4997
0.75	.3011	.2734	2.10	.0440	.4821	3.45	.0010	.4997
0.80	.2897	.2881	2.15	.0395	.4842	3.50	.0009	.4998
0 85	.2780	.3023	2.20	.0355	.4861	3.55	.0007	.4998
0.90	.2661	.3159	2.25	.0317	.4878	3.60	.0006	.4998
0 95	.2541	.3289	2.30	.0283	.4893	3.65	.0005	.4999
1.00	.2420	.3413	2.35	.0252	.4906	3.70	.0004	.4999
1.05	.2299	.3531	2.40	.0224	.4918	3.75	.0004	.4999
1.10	.2179	.3643	2.45	.0198	.4929	3.80	.0003	.4999
1.15	.2059	.3749	2 50	.0175	.4938	3.85	.0002	.4999
1.20	.1942	.3849	2.55	.0154	.4946	3.90	.0002	.5000
1.25	.1826	.3944	2.60	.0136	.4953	3.95	.0002	.5000
1.30	.1714	.4032	2.65	.0119	.4960	4.00	.0001	.5000

TABLE OF CONSTANTS TO BE USED IN FORMULAS FOR PARAMETERS OF PARABOLIC CURVES *

n	K_1	K_2	K_3	K_4	K_5	K_6	K_7	K_8	K_9	K_{10}	n
6	35	101	112	35		8	0.000 108	2.295 040	0.000 960	0.000 240	6
7	56	140	252	48	12		0.000 360	4.990 476	0.005 280	0.002 880	7
8	84	185	504	63		15	0.000 990	9.350 595	0.021 120	0.018 720	8
9	120	236	924	80	20		0.002 376	15.911 111	0.068 640	0.087 360	9
10	165	293	1 584	99		24	0.005 148	25.279 167	0.192 192	0.327 600	10
11	220	356	2 574	120	30		0.010 296	38.133 333	0.480 480	1.048 320	11
12	286	425	4 004	143		35	0.019 305	55.223 611	1.098 240	2.970 240	12
13	364	500	6 006	168	42		0.034 320	77.371 429	2.333 760	7.637 760	13
14	455	581	8 736	195		48	0.058 344	105.469 643	4.667 520	18.139 680	14
15	560	668	12 376	224	56		0.095 472	140.482 540	8.868 288	40.310 400	15
16	680	761	17 136	255		63	0.151 164	183.445 833	16.124 160	84.651 840	16
17	816	860	23 256	288	72		0.232 560	235.466 667	28.217 280	169.303 680	17
18	969	965	31 008	323		80	0.348 840	297.723 612	47.752 320	324.498 720	18
19	1 140	1 076	40 698	360	90		0.511 632	371.466 667	78.450 240	599.074 560	19
20	1 330	1 193	52 668	399		99	0.735 471	458.017 262	125.520 384	1 069.776 000	20
21	1 540	1 316	67 298	440	110		1.038 312	558.768 254	196.125 600	1 854.278 400	21
22	1 771	1 445	85 008	483		120	1.442 100	675.183 930	299.956 800	3 129.094 800	22
23	2 024	1 580	106 260	528	132		1.973 400	808.800 000	449.935 200	5 153.803 200	23
24	2 300	1 721	131 560	575		143	2.664 090	961.223 611	663.062 400	8 303.349 600	24
25	2 600	1 868	161 460	624	156		3.552 120	1 134.133 3	961.440 480	13 110.552 000	25
26	2 925	2 021	196 560	675		168	4.682 340	1 329.279 2	1 373.486 4	20 321.355 600	26
27	3 276	2 180	237 510	728	182		6.107 400	1 548.482 5	1 935.367 2	30 965.875 200	27
28	3 654	2 345	285 012	783		195	7.888 725	1 793.636 3	2 692.684 8	46 448.812 800	28
29	4 060	2 516	339 822	840	210		10.097 568	2 066.704 8	3 702.441 6	68 663.462 240	29
30	4 495	2 693	402 752	899		224	12.816 144	2 369.723 6	5 035.320 6	100 134.216 00	30
31	4 960	2 876	474 672	960	240		16.138 848	2 704.800 0	6 778.316 2	144 193.271 04	31
32	5 456	3 065	556 512	1 023		255	20.173 560	3 074.112 5	9 037.754 9	205 198.116 48	32
33	5 984	3 260	649 264	1 088	272		25.043 040	3 479.911 1	11 942.748	288 797.349 12	33
34	6 545	3 461	753 984	1 155		288	30.886 416	3 924.517 3	15 649.117	402 253.450 56	34
35	7 140	3 668	871 794	1 224	306		37.860 768	4 410.323 8	20 343.853	554 832.345 60	35
36	7 770	3 881	1 003 884	1 295		323	46.142 811	4 939.795 0	26 250.132	758 270.872 32	36
37	8 436	4 100	1 151 514	1 368	342		55.930 680	5 515.466 7	33 632.982	1 027 334.730	37
38	9 139	4 325	1 316 016	1 443		360	67.445 820	6 139.945 8	42 805.614	1 380 481.044	38
39	9 880	4 556	1 498 796	1 520	380		80.934 984	6 815.911 1	54 136 512	1 840 641.392	39
40	10 660	4 793	1 701 336	1 599		399	96.672 342	7 546.112 5	68 057.329	2 436 143.018	40

41	11 480	5 036	1 925 196	1 680	420		114.961 704	8 333.371 4	85 071.661	3 201 787.967	41
42	12 341	5 285	2 172 016	1 763		440	136.138 860	9 180.580 8	105 764.768	4 180 112.068	42
43	13 244	5 540	2 443 518	1 848	462		160.574 040	10 090.705	130 814 318	5 422 848.088	43
44	14 190	5 801	2 741 508	1 935		483	188.674 497	11 066 779	161 002 237	6 992 619.903	44
45	15 180	6 068	3 067 878	2 024	506		220.887 216	12 111.911	197 227.741	8 964 897.312	45
46	16 215	6 341	3 424 608	2 115		528	257.701 752	13 229.279	240 521.635	11 430 244.073	46
47	17 296	6 620	3 813 768	2 208	552		299.653 200	14 422.133	292 061.986	14 496 894.922	47
48	18 424	6 905	4 237 520	2 303		575	347.325 300	15 693.795	353 191.238	18 293 700.734	48
49	19 600	7 196	4 698 120	2 400	600		401.353 680	17 047.657	425 434 901	22 973 484.643	49
50	20 825	7 493	5 197 920	2 499		624	462.429 240	18 487.184	510 521.881	28 716 855.804	50
51	22 100	7 796	5 739 370	2 600	650		531.301 680	20 015.911	610 406.597	35 736 531.667	51
52	23 426	8 105	6 325 020	2 703		675	608.783 175	21 637.446	727 292.966	44 282 224.022	52
53	24 804	8 420	6 957 522	2 808	702		695.752 200	23 355.467	863 660.398	54 646 148.794	53
54	26 235	8 741	7 639 632	2 915		728	793.157 508	25 173.724	1 022 291.9	67 169 224.559	54
55	27 720	9 068	8 374 212	3 024	756		902.022 264	27 096.038	1 206 304.4	82 248 030.072	55
56	29 260	9 401	9 164 232	3 135		783	1 023.448 3	29 126.303	1 419 181.7	100 342 596.69	56
57	30 856	9 740	10 012 772	3 248	812		1 158.620 8	31 268.483	1 664 809.3	121 985 117.54	57
58	32 509	10 085	10 923 024	3 363		840	1 308.812 3	33 526.613	1 947 512.8	147 789 661.64	58
59	34 220	10 436	11 898 294	3 480	870		1 475.388 5	35 904.800	2 272 098.2	178 462 987.64	59
60	35 990	10 793	12 942 004	3 599		899	1 659.812 0	38 407.224	2 643 896.1	214 816 559.19	60
61	37 820	11 156	14 057 694	3 720	930		1 863.648 6	41 038.133	3 068 808.0	257 779 871.03	61
62	39.711	11 525	15 249 024	3 843		960	2 088.571 7	43 801.851	3 553 356.6	308 415 202.84	62
63	41 664	11 900	16 519 776	3 968	992		2 336.368 3	46 702.768	4 104 739.5	367 933 926 20	63
64	43 680	12 281	17 873 856	4 095		1 023	2 608.944 6	49 745.351	4 730 886.3	437 714 498.41	64
65	45 760	12 668	19 315 296	4 224	1 056		2 908.331 7	52 934.133	5 440 519.2	519 322 286.25	65
66	47 905	13 061	20 848 256	4 355		1 088	3 236.691 7	56 273.724	6 243 218.7	614 531 372 06	66
67	50 116	13 460	22 477 026	4 488	1 122		3 596.324 2	59 768.800	7 149 492.4	725 348 504.72	67
68	52 394	13 865	24 206 028	4 623		1 155	3 989.672 1	63 424.113	8 170 848.5	854 039 368.47	68
69	54 740	14 276	26 039 818	4 760	1 190		4 419.329 1	67 244.483	9 319 874.1	1 003 157 353.	69
70	57 155	14 693	27 983 088	4 899		1 224	4 888.045 8	71 234.803	10 610 318.	1 175 575 024.	70
71	59 640	15 116	30 040 668	5 040	1 260		5 398.737 2	75 400.038	12 057 180.	1 374 518 489.	71
72	62 196	15 545	32 217 528	5 183		1 295	5 954.489 6	79 745.224	13 676 801.	1 603 604 904.	72
73	64 824	15 980	34 518 780	5 328	1 332		6 558.568 2	84 275.467	15 486 966.	1 866 883 321.	73
74	67 525	16 421	36 949 680	5 475		1 368	7 214.425 0	88 995.946	17 507 005.	2 168 879 152.	74
75	70 300	16 868	39 515 630	5 624	1 406		7 925.706 4	93 911.911	19 757 905.	2 514 642 495.	75
76	73 150	17 321	42 222 180	5 775		1 443	8 696.261 1	99 028.684	22 262 429.	2 909 800 602.	76
77	76 076	17 780	45 075 030	5 928	1 482		9 530.149 2	104 351.657	25 045 232.	3 360 614 780.	77
78	79 079	18 245	48 080 032	6 083		1 520	10 431.650	109 886.295	28 133 000.	3 874 042 038.	78
79	82 160	18 716	51 243 192	6 240	1 560		11 405.270	115 638.133	31 554 582.	4 457 801 797.	79
80	85 320	19 193	54 570 672	6 399		1 599	12 455.756	121 612.779	35 341 131.	5 120 448 010.	80

* Reprinted from *Journal of American Statistical Association*, December, 1932, by permission.

BIBLIOGRAPHY

ADAMS, THOMAS S. (ed.), "Manual of Charting," Prentice-Hall, Inc., New York, 1924.

BABSON, ROGER W., "Business Barometers for Anticipating Conditions," 21st ed., Babson's Statistical Organization Incorporated, Babson Park, 1929, 373 pages.

"Barlow's Tables of Square Cubes, Square Roots, Cube Roots, and Reciprocals," edited by L. J. COMRIE, Spon and Chamberlain, New York, 1930.

BERRIDGE, WM. A., "Cycles of Unemployment," Houghton Mifflin Company, Boston, 1923, 88 pages.

BEVERIDGE, W. H., "Unemployment: A Problem of Industry," Longmans, Green and Company, New York, 1930, 541 pages.

BOWLEY, ARTHUR L., "Elements of Statistics," 5th ed., Charles Scribner's Sons, New York, 1926 (one-volume edition), 462 pages.

BROWN, T. H., "Problems in Business Statistics," McGraw-Hill Book Company, Inc., New York, 1931.

BURGESS, ROBERT W., "Introduction to the Mathematics of Statistics," Houghton Mifflin Company, New York, 1927.

CHADDOCK, R. E., "Principles and Methods of Statistics," Houghton Mifflin Company, Boston, 1925.

CHADDOCK, R. E., and CROXTON, F. E., "Exercise in Statistical Methods," Houghton Mifflin Company, Boston, 1928.

CONNOR, L. R., "Statistics in Theory and Practice," Isaac Pitman and Sons, New York, 1933.

CRUM, W. L., and PATTON, A. C., "Economic Statistics," A. W. Shaw Company, New York, 1928.

DAY, EDMUND E., "Statistical Analysis," The Macmillan Company, 1925.

DITTMER, CLARENCE G., "Introduction to Social Statistics," A. W. Shaw Company, New York, 1926.

DUBLIN, LOUIS I., "Population Problems," Houghton Mifflin Company, Boston, 1926, 318 pages.

DUNLAP, J. W., and KURTZ, A. K., "Handbook of Statistical Nomographs, Tables, and Formulas," World Book Company, Yonkers-on-Hudson, 1932.

EDIE, LIONEL D. (ed.), "The Stabilization of Business," The Macmillan Company, New York, 1923, 400 pages.

EVANS, G. C., "Mathematical Introduction to Economics," McGraw-Hill Book Company, Inc., New York, 1930.

EZEKIEL, MORDECAI, "Methods of Correlation Analysis," John Wiley & Sons, Inc., New York, 1930.

FISHER, A., "An Elementary Treatise on Frequency Curves," The Macmillan Company, New York, 1923, 240 pages.

"The Mathematical Theory of Probabilities and Its Application to Frequency Curves and Statistical Methods," The Macmillan Company, New York, 1922, 289 pages.

FISHER, IRVING, "Mathematical Investigation in the Theory of Value and Prices," Yale University Press, New Haven, 1926.

FISHER, R. A., "Statistical Methods for Research Workers," Oliver and Boyd, London, 1928.

FLORENCE, P. SARGANT, "The Statistical Method in Economics and Political Science," Harcourt, Brace and Company, New York, 1929.

GAY, EDWIN F., WESLEY C. MITCHELL and others, "Recent Economic Changes," National Bureau of Economic Research, Inc., New York, 1929, 950 pages (2 volumes).

GLOVER, J. W., "Tables of Applied Mathematics, Finance, Insurance, Statistics," 2d ed., George Wahr, Ann Arbor, Mich., 1923, 678 pages.

HABERLER, GOTTFRIED, "Der Sinn der Indexzahlen," J. C. B. Mohr, Tübingen, 1927.

HANEY, LEWIS H., "Business Forecasting," Ginn and Company, Boston, 1931, 373 pages.

HANSEN, A. H., "Business Cycle Theory," Ginn and Company, Boston, 1927.

HARDY, CHARLES O., "Risk and Risk Bearing," The University of Chicago Press, Chicago, 1923, 400 pages.

HARDY, CHARLES O., and COX, GARFIELD V., "Forecasting Business Conditions," The Macmillan Company, New York, 1927.

HARPER, F. H., "Elements of Practical Statistics," The Macmillan Company, New York, 1930.

HASKEL, ALLAN C., "Graphic Charts in Business," Codex Book Company, New York, 1922.

HODGMAN, C. D., "Mathematical Tables" (from "Handbook of Chemistry and Physics"), Chemical Rubber Publishing Company, Cleveland, 1931.

JEROME, HARRY, "Migration and Business Cycles," National Bureau of Economic Research, Inc., New York, 1926, 256 pages.
"Statistical Method," Harper and Brothers, New York, 1924.

JORDAN, DAVID F., "Practical Business Forecasting," Prentice-Hall, Inc., New York, 1927.

Journal of the American Statistical Association, published quarterly by the American Statistical Association, New York.

KARSTEN, K. G., "Charts and Graphs," Prentice-Hall, Inc., New York, 1925.

KELLEY, TRUMAN, L., "Statistical Method," The Macmillan Company, New York, 1923.

KING, WILLFORD I., "Employment, Hours, and Earnings in Prosperity and Depression," National Bureau of Economic Research, Inc., New York, 1923, 147 pages.
"Index Numbers Elucidated," Longmans, Green and Company, New York, 1930.
"The National Income and Its Purchasing Power," National Bureau of Economic Research, Inc., New York, 1930.

KUZNETS, SIMON S., "Secular Movements in Production and Prices," Houghton Mifflin Company, Boston, 1930.
"Seasonal Variations in Industry and Trade," National Bureau of Economic Research, New York, 1933.

LACROIX and RAGOT, "A Graphic Table Combining Logarithms and Antilogarithms," The Macmillan Company, New York, 1927.

LOVITT, W. V., and HOLTZCLAW, H. F., "Statistics," Prentice-Hall, Inc., New York, 1929.

MACAULAY, F. R., "The Smoothing of Time Series," National Bureau of Economic Research, Inc., New York, 1931.

MCMILLAN, A. W., "Measurement in Social Work," University of Chicago Press, Chicago, 1930.

MILLS, FREDERICK C., "Statistical Methods Applied to Economics and Business," Henry Holt and Company, New York, 1924, 604 pages.

"The Behavior of Prices," National Bureau of Economic Research, New York, 1927.

MILLS, F. C., and DAVENPORT, D. H., "A Manual of Problems and Tables in Statistics with Notes on Statistical Procedure," Henry Holt and Company, New York, 1925, 203 pages.

MITCHELL, WESLEY C., "Business Cycles, The Problem and Its Setting," National Bureau of Economic Research, Inc., New York, 1927.

MOORE, HENRY LUDWELL, "Economic Cycles: Their Law and Cause," The Macmillan Company, New York, 1914, 149 pages.

"Generating Economic Cycles," The Macmillan Company, New York, 1923, 141 pages.

"Synthetic Economics," The Macmillan Company, New York, 1929.

MUDGETT, BRUCE D., "Statistical Tables and Graphs," Houghton Mifflin Company, Boston, 1930.

OLIVIER, MAURICE, "Les nombres indices de la variation des prix," Marcel Giard, Paris, 1927.

PASSANO, L. M., "Calculus and Graphs," The Macmillan Company, New York, 1932.

PEARSON, KARL, "Tables for Statisticians and Biometricians," University Press, Cambridge, 1914, 144 pages. (Out of print.)

PEIRCE, B. O., "A Short Table of Integrals," Ginn and Company, Boston, 1910.

PERSONS, W. M., "The Construction of Index Numbers," Houghton Mifflin Company, Boston, 1928.

"Forecasting Business Cycles," John Wiley and Sons, Inc., New York, 1931.

"Recent Economic Changes," Report of the Committee on Recent Economic Changes of the President's Conference on Unemployment, McGraw-Hill Book Company, Inc., New York, 1929. Two volumes.

REINHARDT, J. M., and DAVIES, G. R., "Principles and Methods of Sociology," Prentice-Hall, Inc., New York, 1932.

RICE, S. A., "Methods in Social Science," University of Chicago Press, Chicago, 1931.

(ed.), "Statistics in Social Studies," University of Pennsylvania Press, Philadelphia, 1930.

RIEGEL, ROBERT, "Elements of Business Statistics," rev. ed., D. Appleton and Company, New York, 1927, 549 pages.

RIETZ, H. L. (ed.), "Handbook of Mathematical Statistics," Houghton Mifflin Company, Boston, 1924.

RIGGLEMAN, J. R., and FRISBEE, I. N., "Business Statistics," McGraw-Hill Book Company, Inc., New York, 1932.

SCHLUTER, W. C., "How to Do Research Work," Prentice-Hall, Inc., New York, 1927, 137 pages.

SCHULTZ, HENRY, "Statistical Laws of Demand and Supply," University of Chicago Press, Chicago, 1928.

SECRIST, HORACE, "An Introduction to Statistical Methods," The Macmillan Company, New York, 1925, 584 pages.

"Banking Ratios," Claremont Colleges Research Studies I, Stanford University Press, 1930, 608 pages.

SNIDER, JOSEPH L., "Business Statistics," 2d ed., McGraw-Hill Book Company, Inc., New York, 1932.

SNYDER, CARL, "Business Cycles and Business Measurements," The Macmillan Company, New York, 1927.

STOCKWELL, H. G., "How to Read a Financial Statement," The Ronald Press Company, New York, 1925, 443 pages.

"How to Read a Profit and Loss Statement," The Ronald Press Company, New York, 1927, 411 pages.

Survey of Current Business, Published monthly by the United States Department of Commerce, Washington, D. C.

SUTCLIFFE, WILLIAM G., "Elementary Statistical Methods," McGraw-Hill Book Company, Inc., New York, 1925, 338 pages.

"Statistics for the Business Man," Harper and Brothers, New York, 1930, 243 pages.

THOMAS, DOROTHY S., "Social Aspects of the Business Cycle," A. A. Knopf, New York, 1927.

THURSTONE, L. L., "Fundamentals of Statistics," The Macmillan Company, New York, 1925, 237 pages.

VANDERBLUE, HORACE N., "Problems in Business Economics," A. W. Shaw Company, New York, 1929.

WAGEMANN, ERNST, "Economic Rhythm," McGraw-Hill Book Company, Inc., New York, 1930.

WALKER, HELEN M., "Studies in the History of Statistical Method," The Williams & Wilkins Company, Baltimore, 1929, 229 pages.

WALSH, C. M., "The Measurement of General Exchange Value," The Macmillan Company, New York, 1901.

WHITE, R. CLYDE, "Social Statistics," Harper and Brothers, New York, 1933, 471 pages.

WOLFENDEN, H. H., "Population Statistics," Actuarial Society of America, New York, 1925.

YOUNG, BENJAMIN F., "Statistics as Applied in Business," The Ronald Press Company, New York, 1925, 639 pages.

YULE, G. UDNY, "An Introduction to the Theory of Statistics," 9th rev., J. B. Lippincott Company, Philadelphia, 1929, 424 pages.

INDEX